Title

Evidence for negative selection against somatic mutations in fibroblasts with high N-ethyl-N-nitrosourea-induced mutation load

Short Title

Mutation accumulation and negative selection

Authors

Ronald Cutler[1,2,=,*], Johanna Heid[1,=], Shixiang Sun[1], Moonsook Lee[1], Alexander Y. Maslov[1,3], Lei Zhang[4], Simone Sidoli[2], Xiao Dong[4, 5], Jan Vijg[1,*]

Affiliations

[1] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[2] Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[3] Laboratory of Applied Genomic Technologies, Voronezh State University of Engineering Technology, Voronezh 394000, Russia

[4] Institute on the Biology of Aging and Metabolism, University of Minnesota, Minneapolis, MN 55455, USA

[5] Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455, USA

[=] Authors contributed equally to this work

[*] Correspondence to J.V. (jan.vijg@einsteinmed.edu) & R.C. (ronald.cutler@einsteinmed.edu)

Abstract

Advanced sequencing has revealed that thousands of mutations accumulate with age in most human tissues. While some can clonally expand and cause disease, the maximum mutation load a cell can tolerate without functional decline is unknown. We addressed this by repeatedly treating proliferating human primary fibroblasts with a low dose of the mutagen N-ethyl-N-nitrosourea and quantifying somatic mutation burden using single-cell whole-genome sequencing. Mutation load increased linearly to ~56,000 single-nucleotide variants per cell with only a slight reduction in growth rate. Analysis showed negative selection against potentially deleterious mutations in coding and non-coding regions, including sequences tied to pathways essential for cell growth and identity. This selective removal of harmful variants likely enables cells to maintain growth functions despite extreme mutation burden. Because most adult tissues are non-dividing and cannot benefit from negative selection based on growth, mutations that accumulate during aging may have pronounced functional consequences.

46 # Introduction

47 Aging is a complex process of continuous decline in cellular and tissue function that is
48 accompanied with increased disease risk(*1*). DNA damage is thought to play a central role in this
49 process and may be a universal cause of aging(*2*). Among the major molecular consequences of
50 DNA damage during aging are somatic DNA mutations. Somatic mutations are the result of errors
51 during repair or replication of a damaged DNA template and can only be purged via the death of
52 the cell or organism. The accumulation of somatic mutations has since long been implicated as a
53 major cause of aging(*3, 4*), which is strongly supported by the age-related exponential increase of
54 cancer, a disease known to be caused by mutations(*5*). Somatic mutations are difficult to analyze
55 because of their random nature, with very low variant allele frequencies indistinguishable from
56 technical artifacts by traditional bulk-sequencing methods. With the recent emergence of single-
57 cell and single-molecule assays, this situation has changed and we are now able to quantitatively
58 analyze most small mutations, i.e. base substitutions and small insertions and deletions, at high
59 accuracy. Application of such methods has provided ample evidence that somatic mutation rate is
60 far higher than the germline mutation rate and that thousands of somatic mutations accumulate
61 with age in most, if not all, human tissues(*6-10*), including postmitotic tissues such as the brain
62 and heart(*11, 12*). In addition, evidence has emerged that postzygotic, somatic mutations are a
63 cause of a wide variety of human diseases other than cancer(*13, 14*). Like cancer, many of these
64 age-related effects may be due to clonally amplified mutations(*15-17*). Indeed, in blood, somatic
65 mutations lead to clonal hematopoiesis, which has been associated with increased risk of age-
66 related diseases, such as cardiovascular disease, as well as overall mortality(*18*).

67 What thus far remains unknown is the possible functional impact, collectively, of thousands of
68 random mutations accumulating in a non-cancerous cell, for example, by adversely affecting gene-

69 coding or gene-regulatory sequences relevant for the given cell type. If so, one would expect a
70 limit to the burden of random mutations a cell can tolerate or selection against mutations in
71 important functional regions. Here, we directly tested this hypothesis by treating actively
72 proliferating human primary fibroblasts with a low, repeated dose of N-ethyl-N-nitrosourea
73 (ENU), a powerful mutagen that induces almost exclusively single-nucleotide variants (SNVs).
74 Our results show that 9 cycles of ENU treatment over the course of ~24 population doublings
75 linearly increased mutation frequency up to ~56,000 SNVs per cell, with only subtle effects on
76 cell growth and death rates. We show that this tolerance for somatic mutations could be explained
77 by negative selection against potentially deleterious SNVs in coding and non-coding genomic loci
78 as well as selection against SNVs across the functional genomic loci associated with cell growth
79 and cell identity. These results suggest that accumulating somatic mutations may have adverse
80 effects on cellular function when negative selection is not possible, as is the case in most cells and
81 tissues in adult organisms, which are not mitotically active(*19*).

82

83 # Results

84 ## Repeated Mutagen Treatment of Normal Cells Only Slightly Affects Cell
85 Growth and Death

86 To determine the impact of increased somatic mutation burden on primary human cells *in vitro*,
87 we treated low-passage fetal lung fibroblasts (IMR-90) multiple times with a sublethal dose of 50
88 μg/mL of ENU. Each treatment was followed by a recovery period of 7 days, after which 1 million
89 cells (determined by cell counting) were replated and treated again. This process was repeated for
90 9 cycles (Figure 1A; Methods). To monitor cell growth and toxicity, we measured cell number and

91 apoptosis rate at 3 and 7 days after each treatment to assess both short- and long-term effects of
92 the treatment (Methods). We measured cell growth by calculating population doublings (PDs), and
93 found that repeated treatment with ENU gradually decreased cell growth rate over the course of
94 the experiment, as at cycle 9 there were a total of 28.25±1.87 (mean ± standard deviation) PDs for
95 control cells compared to 24.1±1.34 PDs for ENU-treated cells (Figure 1B). The effect of ENU
96 treatment on cell growth was more pronounced at the 3 day time point (Figure 1C) compared to
97 the 7 day time point (Figure 1D), which mostly explains the decrease in overall growth rate (Figure
98 1B). A larger effect on growth rate at the 3 day timepoint was expected due to the DNA damage
99 response that is known to occur early and then subsides over time. To identify the contribution of
100 cell death to the slight decrease in growth rate due to mutagen exposure, we measured markers of
101 early and late apoptosis at the 3 and 7 day time points. The results showed a slight, but consistent
102 increase in early and late apoptosis in the ENU-treated group as compared to the control cells at
103 both time points (Figures 1E-F and 1G-H).

104

105

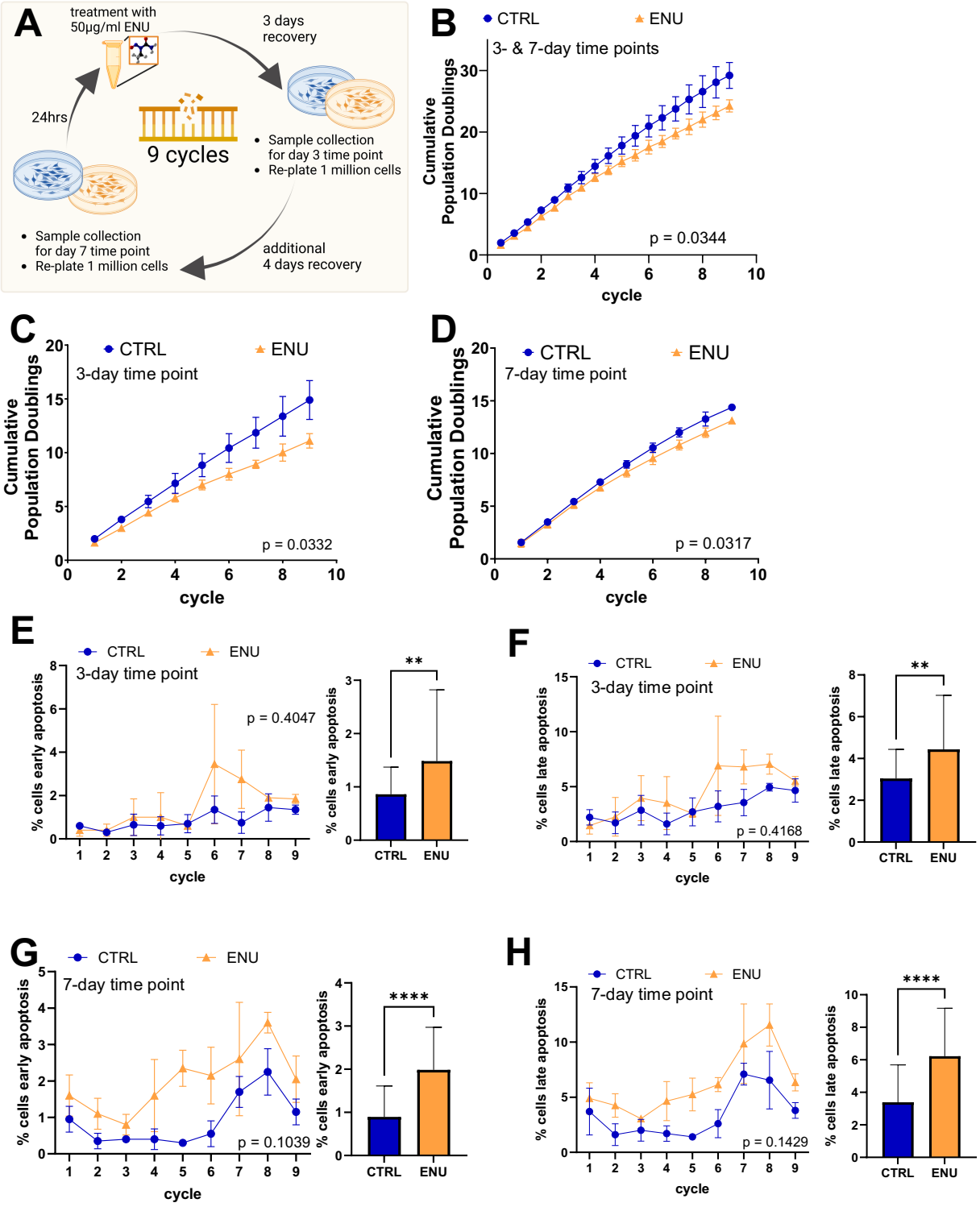

A
treatment with 50µg/ml ENU
3 days recovery
24hrs
9 cycles
Sample collection for day 3 time point
Re-plate 1 million cells
Sample collection for day 7 time point
Re-plate 1 million cells
additional 4 days recovery
B
CTRL
ENU
3- & 7-day time points
Cumulative Population Doublings
p = 0.0344
cycle
C
3-day time point
p = 0.0332
D
7-day time point
p = 0.0317
E
% cells early apoptosis
p = 0.4047
**
F
% cells late apoptosis
p = 0.4168
**
G
p = 0.1039
****
H
p = 0.1429
****

Figure 1. Repeated Mutagen Treatment of Normal Cells Only Slightly Affects Cell Growth and Death

(A) *Schematic depiction of the experimental design. Human fetal lung fibroblasts (IMR-90) were treated with a sublethal dose of 50 µg/mL of N-ethyl-N-nitrosourea (ENU), which was followed by a recovery period of 7 days. At 3-day and 7-day time points, 1 million cells were passaged and a sample was taken for further analysis. This process was repeated for 9 cycles (Methods).* ***(B)*** *Cumulative population doublings (PDs) of the control (CTRL) and ENU-treated (ENU) groups throughout the experiment. Statistics: n=3; Data represent mean ± S.D.; P-value estimated using a two-way repeated measures ANOVA.* ***(C)*** *Cumulative PDs of the CTRL and ENU groups assessed at the 3-day time point. Statistics: Same as in (B).* ***(D)*** *Same as in (C), but at 7-day time point. Statistics: Same as in (B).* ***(E)*** *Percentage of cells detected in early apoptosis at the 3-day time point at each cycle (left) or with all cycles pooled (right). Statistics: n=2; Data represent mean ± S.D.; P-value of data from each cycle was estimated using a two-way repeated measures ANOVA. P-value of data from all cycles pooled was estimated using a paired t-test; **: p <= 0.005; ****: p<= 0.0001.* ***(F)*** *Same as in (E), but for late apoptosis. Statistics: Same as in (E).* ***(G)*** *Same as in (E), but for 7-day time point. Statistics: Same as in (E).* ***(H)*** *Same as in F, but for 7-day time point. Statistics: Same as in (E).*

To analyze possible effects of a DNA damage response induced by ENU treatment in more detail, we performed bulk mRNA-sequencing analysis on samples collected at cycles 1 and 9 for the control group and cycle 9 for the ENU-treated group (Figures S1A-K; Table S1; Methods). As a positive control for the effect of cell passsaging, we first compared the control cycle 9 and the control cycle 1 groups, which resulted in 979 significantly differentially expressed genes (DEGs) (Figures 2A; Table S2)(*20*). We then performed targeted gene set enrichment analysis (GSEA) which showed that cell cycle-related genes were significantly decreased, while apoptosis-related genes were not significantly changed (Figures 2B-C). As cells are known to become senescent during passaging, we also tested for an enrichment of senescence-related genes using the SenMayo gene set(*21*), which were found to be significantly increased, as expected (Figure 2D).

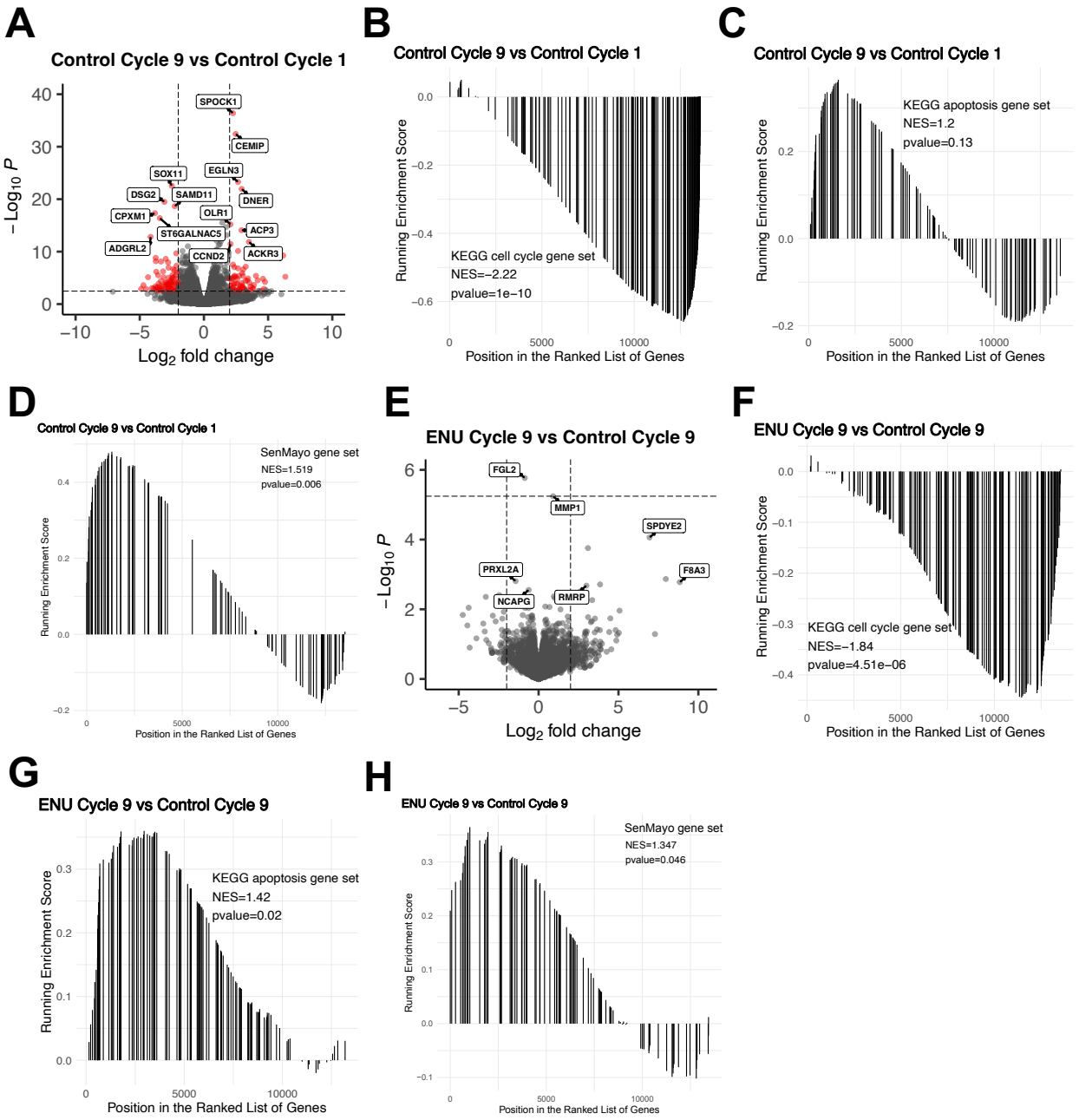
134 Furthermore, we found the DEGs to be positively correlated with DEGs found in a previous study
135 on replicative senescence (Figure S1N).

136 To assess the effects of the repeated ENU treatment, we compared the ENU-treated cycle 9 and
137 the control cycle 9 groups. Unexpectedly, this resulted in only 2 significant DEGs (Figure 2E;
138 Table S2; Supplementary Note 1). However, after performing GSEA (which does not rely on p-
139 value thresholds), we found that repeated ENU treatment had a similar, albeit weaker, effect as
140 cell passaging – a significant decrease of cell cycle-related genes and a significant increase in
141 senescence-related genes (Figure 2F, 2H). In contrast to the effect of cell passaging, we found a
142 significant increase in the abundance of apoptosis-related genes in the ENU-treated cells relative
143 to the controls (Figure 2G), confirming the results of the direct apoptosis measurements (Figures
144 1G-H).

145

146

147


A
Control Cycle 9 vs Control Cycle 1
SPOCK1
CEMIP
EGLN3
SOX11
DSG2
SAMD11
DNER
CPXM1
OLR1
ACP3
ST6GALNAC5
ADGRL2
CCND2
ACKR3
−Log10 P
Log2 fold change
B
Control Cycle 9 vs Control Cycle 1
KEGG cell cycle gene set
NES=−2.22
pvalue=1e−10
Running Enrichment Score
Position in the Ranked List of Genes
C
Control Cycle 9 vs Control Cycle 1
KEGG apoptosis gene set
NES=1.2
pvalue=0.13
Running Enrichment Score
Position in the Ranked List of Genes
D
Control Cycle 9 vs Control Cycle 1
SenMayo gene set
NES=1.519
pvalue=0.006
Running Enrichment Score
Position in the Ranked List of Genes
E
ENU Cycle 9 vs Control Cycle 9
FGL2
MMP1
SPDYE2
F8A3
PRXL2A
NCAPG
RMRP
−Log10 P
Log2 fold change
F
ENU Cycle 9 vs Control Cycle 9
KEGG cell cycle gene set
NES=−1.84
pvalue=4.51e−06
Running Enrichment Score
Position in the Ranked List of Genes
G
ENU Cycle 9 vs Control Cycle 9
KEGG apoptosis gene set
NES=1.42
pvalue=0.02
Running Enrichment Score
Position in the Ranked List of Genes
H
ENU Cycle 9 vs Control Cycle 9
SenMayo gene set
NES=1.347
pvalue=0.046
Running Enrichment Score
Position in the Ranked List of Genes

148 **Figure 2. RNA-Sequencing Reveals Increased Cell Cycle Inhibition and Apoptosis in ENU-**
149 **treated Cells**

150 ***(A)*** *Differential gene expression (DGE) of control cycle 9 vs. control cycle 1 groups. Red points*
151 *indicate* Log_2 *fold change>2 and adjusted p-value<0.05. See Table S1 for sample details and*
152 *Table S2 for DGE results. Statistics: n=3; DGE p-values estimated using a Wald test followed by*
153 *Benjamini-Hochberg correction for multiple hypothesis testing.* ***(B-D)*** *Gene set enrichment*
154 *analysis (GSEA) of DGE results from the control cycle 9 vs. control cycle 1 comparison for KEGG*
155 *cell cycle (B), KEGG apoptosis (C), and SenMayo senescence (D) gene sets. NES=normalized*
156 *enrichment score. Statistics: n=3; p-values estimated using a permutation test.* ***(E)*** *Same as in (A),*
157 *but for the ENU-treated cycle 9 vs. control cycle 9 comparison. See Supplementary Note 1 for*
158 *batch correction note. Statistics: Same as in (A).* ***(F-H)*** *Same as in (B-D), but for the ENU-treated*
159 *cycle 9 vs. control cycle 9 comparison. Statistics: Same as in (B-D).*

160

161 Taken together, these phenotypic and molecular data suggest that the observed slight decrease in
162 cell growth rate after repeated ENU treatment is, at least in part, due to DNA damage-inducd cell
163 cycle inhibition and increased apoptosis. As the resulting effect sizes due to repeated ENU
164 treatment (i.e. mutation accumulation) were relatively small, these results demonstrate the
165 tolerence of normal primary cells proliferating in culture to this mutagen, which was previously
166 known to have low cell toxicity when used at low doses(*22*).

167

168 Repeated Mutagen Treatment of Normal Cells Results in Extremely High
169 Mutation Burden

170 Somatic DNA mutations in normal (non-cancerous) cells, which are mostly unique to each cell,
171 cannot be distinguished from sequencing artifacts after bulk whole genome sequencing. Hence,
172 we used our established and highly accurate single-cell whole genome sequencing assay, which
173 has been extensively validated by parallel analysis of kindred single-cell clones(*23, 24*), to analyze
174 mutation burden in 21 single-cells from control and ENU-treated groups collected in 3 independent

175 batches at cycles 3, 6, and 9 from the repeated ENU treatment experiment (Figure 1A, Table S3,
176 Methods). Cells for mutation analysis were collected at the 7-day time point after each treatment
177 cycle, as DNA damage responses should be abated by that time and all mutations fixed (Figure
178 2D). We then first corrected the mutations that were actually observed for the less than 100%
179 genome coverage and sensitivity, which provided estimated mutation burdens higher than the
180 observed burdens (Figures S2A-G; Methods). We also corrected for technical artifacts in the
181 control group for cycle 1, which reduced mutation burden somewhat but had no effect at all on the
182 ENU-induced mutation burden (Figures S3A-D; Methods). We found that the number of estimated
183 SNVs per cell in the control group was initially 1,996±332 at cycle 1, which then increased to
184 19,064±823 following 3 ENU treatment cycles, (~7.8-fold increase compared to cycle 1, p-
185 value=$1.54*10^{-6}$), and then 33,098±1,785 after 6 ENU treatment cycles (~1.7-fold increase
186 compared to cycle 3, p=$1.90*10^{-5}$), and finally reached 55,954±4,066 after the 9th ENU treatment
187 cycle (~1.7-fold increase compared to the cycle 6, p =$3.33*10^{-9}$) (Figures 3A and S2H). T>A and
188 T>C substitutions, known to be induced by ENU(*22*), accounted for a large proportion of SNVs
189 and showed the same trends as when analyzing the total mutation burden (Figures S2K-L).
190 Importantly, there was no significant difference between control cells at cycle 1 and 9, ruling out
191 that the increase in SNVs in ENU-treated cells was due to passaging. We also checked for the
192 possibility of clonal expansion by looking at shared mutations between cells within each group,
193 but only detected 8 SNVs that were shared between 2 cells in the ENU-treated group at cycle 6
194 (Figure S2F). This indicates a shared common ancestor, but not clonal expansion. This was
195 expected as clones should not have the opportunity to grow over the relatively short number of
196 cell divisions and the small number of individual cells analyzed. We conclude that our treatment
197 paradigm increased mutational load to supraphysiological levels without serious toxicity, with the

198 increase in SNVs over the course of the experiment occurring linearly in a cycle-dependent manner
199 (Figure 3B).

200 While ENU is known to mainly induce base substitution mutations through alkylation damage, the
201 estimated number of small insertions and deletions (INDELs) per cell was also observed to
202 increase due to ENU treatment, starting from 123±49 in the control group at cycle 1, to 246±49 in
203 the ENU-treated group at cycle 3 (~1.8-fold increase, $p=4.10*10^{-3}$); however, there was no further
204 increase at cycles 6 and 9 (Figures S2I-J). As with SNVs, there was no significant difference in
205 the amount of estimated INDELs between control cells from cycle 1 and cycle 9. INDELs have
206 been previously observed to be induced by ENU, albeit at low frequencies(*25*), and could be
207 caused by DNA polymerase slippage during DNA replication at sites of alkylation damage. While
208 the lack of an increase in INDELs after cycle 3 in the ENU-treated group may be suggestive of
209 negative selection acting to prevent a further increase, as was seen in a recent study (*26*), our

current finding remains inconclusive due to the low amount of observed INDELs per cell (~20 INDELs per cell).

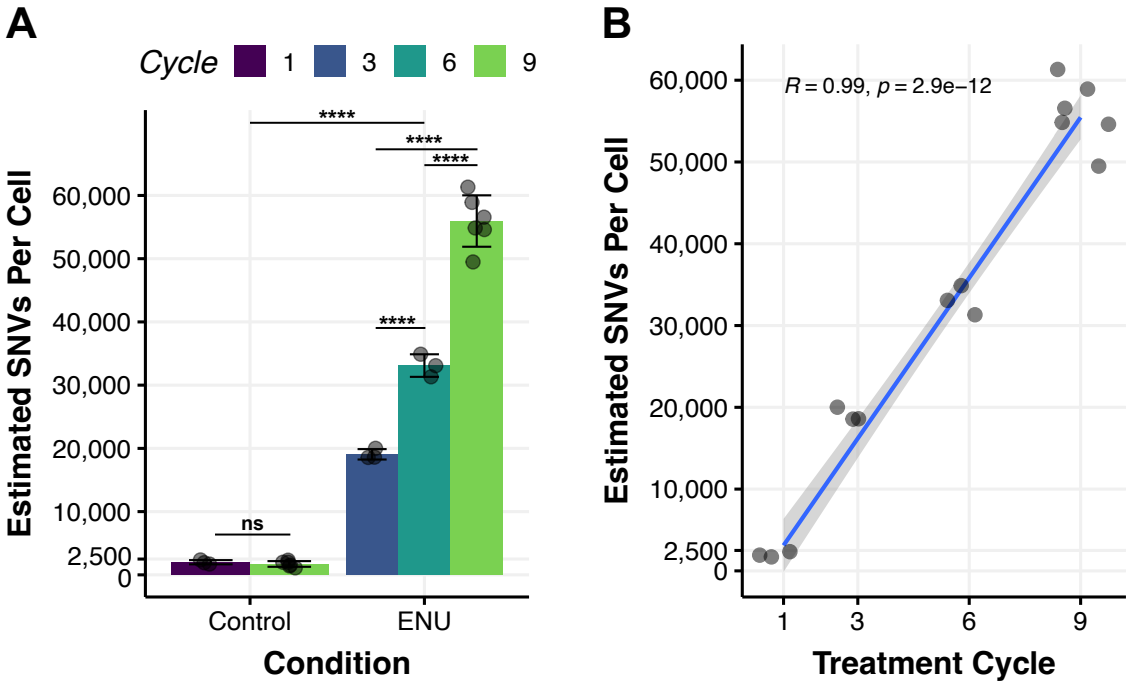


Figure 3. Repeated Mutagen Treatment of Normal Cells Results in Extremely High Mutation Burden.

(A) *Estimated load of somatic single nucleotide variants (SNVs) per cell after correction for genome coverage, sensitivity, and technical artifacts (Methods). See Figure S2H for observed SNV load per cell and Figures S2K-L for ENU-specific mutation burden. See Table S3 for sample details and Table S10 for mutation catalogue. Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D. p-values of comparisons between cycles within the same condition were estimated with a two-sided linear model with estimated marginal means, Tukey's HSD. P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA and shown at very top of plot. p-value legend: ns:p > 0.05, ****: p <= 0.0001.* ***(B)*** *Fit of linear model to estimated SNV load per cell shown in (A) vs. the treatment cycle cells were collected at. Note that cells from control cycle 1 are included and control cycle 9 is omitted. Shown at top is the fit of a linear equation. Statistics: data represent mean ± S.D.; Pearson R correlation coefficient and p-value of the correlation was estimated using a T-test.*

We then characterized the underlying SNV mutational processes in the aggregated mutations from the cells at all cycles within the control and ENU-treated groups (Methods). The ENU-treated

230 group showed an expected increase in the relative contribution of thymine-to-adenine (T>A),
231 followed by thymine-to-cytosine (T>C) base transversions (Figures 4A, S2K, S3A). This is in
232 keeping with previously published ENU-induced mutational spectra by us and others(*22, 27*). We
233 then extracted SNV mutational signatures *de novo* using non-negative matrix factorization with a
234 reference set of mutations from previous single-cell whole genome studies (*23, 28-30*) (Figures
235 S3E-F) and identified two signatures, labeled as SBSA and SBSB, which were dominant in the
236 cells of the ENU-treated and control groups, respectively (Figures 4B and S3G). Importantly, we
237 found that SBSA resembled the SIGNAL human signature associated with ENU-treatment,
238 validating the source of the induced mutations (Figure 4C)(*31*). We then compared our *de novo*
239 signatures to those in the COSMIC database and found that SBSB was most similar to SBS5 a
240 'clock-like' mutational signature that has been previously found to correlate with cellular
241 replication during aging (*28, 32*) (Figure 4D). Though relatively few INDELs were observed (369
242 in total), we also provide a characterization of their mutational spectra (Figure S3H;
243 Supplementary Note 2).

244

245

246

247

248

249

250

251

252

253

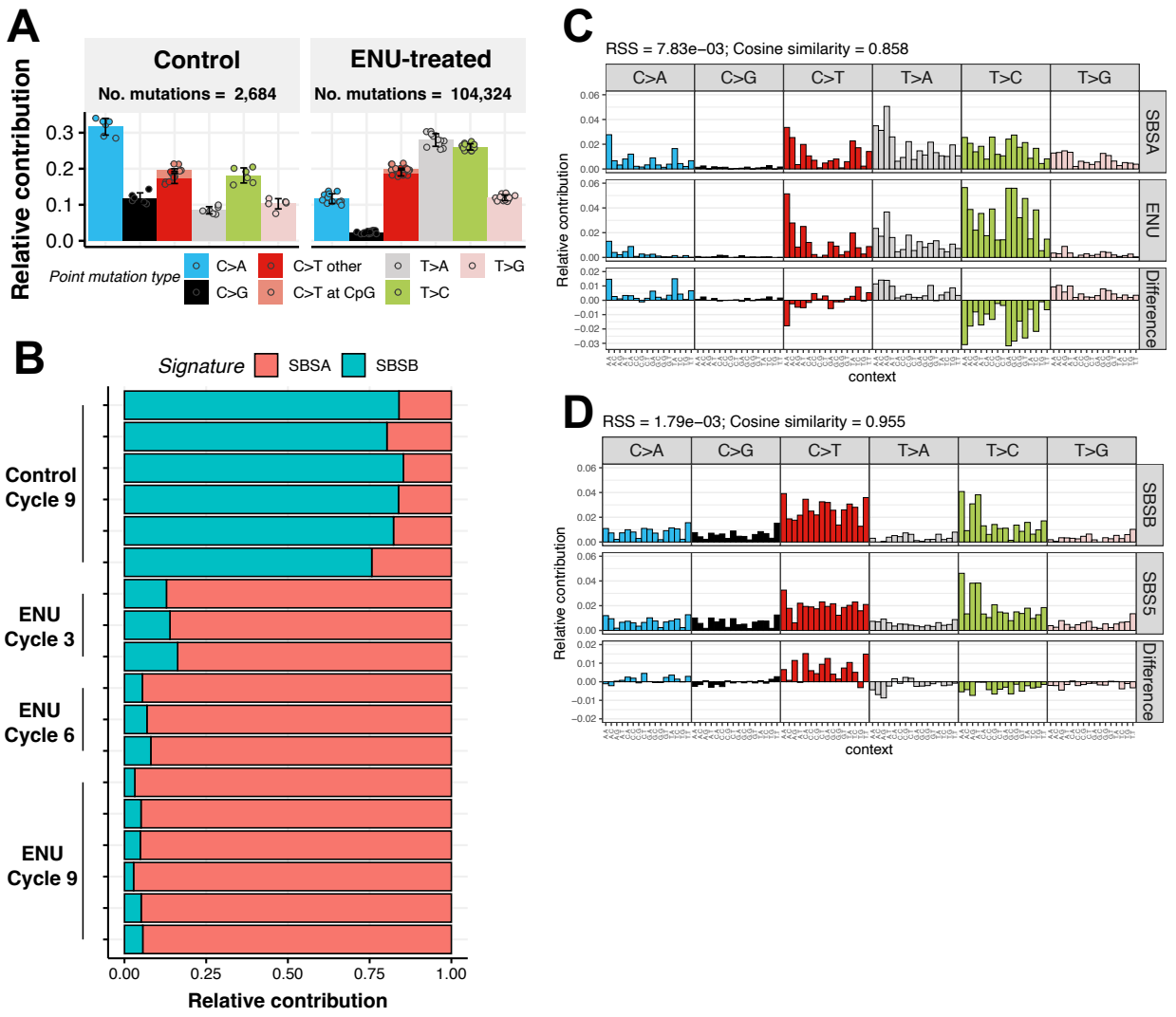


Figure 4. Mutational Spectra And Signatures Resulting From Repeated Mutagen Treatment

(A) *Mutational spectra of the relative contribution of SNV types per cell grouped by condition. See Figure S3A for groups stratified by cycle. Note that cells from the control cycle 1 group were excluded from the mutational spectra and signatures analyses due to abnormally high C>T frequency suggestive of a sample preparation artifact (Figures S3A-D; Methods). Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; Data represent mean ± S.D.* ***(B)*** *Relative contribution of de novo mutational signatures SBSA and SBSB to individual cells (shown as rows) grouped by condition and cycle. See Figure S3F for determination of number of signatures to extract and Figure S3G for absolute contribution.* ***(C)*** *Comparison of de novo mutational signature SBSA to the ENU signature from the SIGNAL database. Statistics: cosine similarity 0.858, residual sum of squares=7.83*10^{-3}.* ***(D)*** *Comparison of de novo mutational signature SBSB to most similar fitted signature SBS5 from the COSMIC database. Statistics: cosine similarity=0.955, residual sum of squares= 1.79*10^{-3}.*

Negative Selection Against ENU-Induced Accumulation of Damaging Coding and Non-Coding Variants

The above results show that the mutation burden of primary cells after 9 cycles of ENU treatment increased to an estimated ~56,000 SNVs per cell (i.e., ~10 times the SNV burden typically found in normal cells from aged tissues(*6*)), with only a slight effect on cell growth and death. From this, we hypothesized that the sustained cell survival and growth in the face of a very high mutation burden could be explained by negative selection against harmful mutations. Our dataset presented a unique opportunity to test this hypothesis as there were a total of 108,600 SNVs observed collectively in all cells. Note that this was smaller than the estimated number after corrections for sensitivity and genome coverage as mentioned above (Methods), but still expected to provide sufficient power for statistical testing of selection pressure (Figure S2H).

To obtain a random background distribution of expected mutations against which to compare the observed mutations in testing for selection pressure, we used the SigProfilerSimulator tool (*33*). This method controls for variation in genomic coverage, mutational signatures, and sequence context of the observed mutations. With this, we simulated 10,000 instances of random mutations for each treatment cycle in both the control and ENU conditions (Methods). Following this, both observed and expected SNVs in coding regions were annotated as synonymous (non-protein-altering) or non-synonymous (protein-altering) variants, using the Ensemble Variant Effect Predictor program (Table S4)(*34*). The simulations of expected mutations performs well in predicting the number of synonymous variants ($R=0.94$, $p=3.6*10^{-10}$; Figure S4A), which should be under minimal selection. We then measured selection pressure within each group by calculating the Log_2(observed/expected) (written as Log_2(O/E) hereafter) ratios for each variant type using the frequency of observed and expected variants (*35*). The direction of selection pressure (i.e. neutral,

292 positive, or negative) was ascertained by comparing the resulting Log_2(O/E) ratio to 0, where no
293 difference from 0 is what would be expected under neutral selection (*36*) (Methods). We then used
294 bootstrapping to determine the threshold for the frequency of observed variants needed to obtain
295 sufficient power for statistically significant results ($p<0.05$) at various Log_2(O/E) ratios (Figures
296 S4B; Table S5; Methods). Finally, to account for the known effect of DNA repair in depleting the
297 frequency of mutations within transcribed regions through transcription-coupled repair (TCR) and
298 mismatch repair (MMR) (*37, 38*), we normalized the Log_2(O/E) ratios of variants within
299 transcribed regions by the Log_2(O/E) ratio of mutations within intronic regions (assumed to be
300 equally transcribed, but under neutral selection; Methods).

301 With this approach, we found that while there were significant cycle-dependent increases in both
302 synonymous and non-synonymous variant frequencies in the ENU-treated group relative to the
303 controls (Figure 5A), the Log_2(O/E) ratios of non-synonymous variants at cycles 6 and 9 in the
304 ENU-treated group indicated that the frequency of these damaging variants was significantly lower
305 than what would be expected, indicating negative selection (Figures 5B and S4C-E, right panels).
306 The negative Log_2(O/E) ratios of non-synonymous variants at cycles 6 and 9 corresponded to a
307 ~20% reduction in mutation burden, equivalent to a decrease of 14.5 ± 9.1 and 19.9 ± 12.5 variants
308 per cell at cycles 6 and 9, respectively (Figure 5C). In regards to synonymous variants, the
309 Log_2(O/E) ratios of these variants trended towards 0 across all cycles in the ENU-treated group,
310 which would suggest neutral selection (as would be expected for these benign variants). However,
311 the frequency of observed synonymous variants was below the threshold set by the previously
312 described power analysis, which prevented us from making any definite conclusions (Figures 5B
313 and S4C-E, left panels; Table S5).

314 We then sought to verify our finding of negative selection against non-synonymous variants by
315 using a method which calculates dN/dS ratios (i.e. the dNdScv program), an alternative measure
316 of selection commonly used in cancer or evolutionary studies (Methods)(*39*). This metric
317 compares the observed number of non-synonymous mutations (missense, nonsense, and splice-
318 site) to the number expected under neutral evolution, using synonymous mutations as a neutral
319 reference. A global dN/dS ratio significantly greater than 1 indicates positive selection, while a
320 ratio less than 1 suggests negative selection. As this is effectively the same as calculating the O/E
321 of non-synonymous mutations over the O/E of synonymous mutations, we calculated the
322 Log_2(O/E) non-synonymous/synonymous (N/S) to facilitate the comparison. Synonymous
323 variants, assumed to be functionally neutral, occur in the same sequence contexts and at the same
324 depth as non-synonymous variants, so they serve as an internal baseline for the true mutation rate
325 introduced by ENU. A deficit of non-synonymous variants relative to this neutral baseline is
326 therefore attributable to negative selection rather than to sequencing bias or differences in
327 mutagenesis. Log_2(O/E) (N/S) ratios were significantly below 0 at cycles 6 and 9 in the ENU-
328 treated group (Figures 5D, S4F-G), indicating negative selection and corroborating our previous
329 result (Figure 5B). Before reviewing the results of the dNdScv method, it should be noted that
330 there are important differences between this method and our O/E method. Unlike our O/E method,
331 which empirically estimates expected mutations via genome-wide simulations accounting for
332 coverage and mutational signatures (*35*), the dNdScv method relies on probabilistic modeling to
333 estimate the expected mutations, assumes uniform genomic coverage, and expects abundant
334 mutations in coding regions. Therefore, the assumptions of this method make it less suitable for
335 our single-cell whole genome sequencing data type, i.e., variable genome coverage and a limited
336 number of coding mutations (Figures S2A and 5A). Expectedly, global dN/dS ratios were in

337 contrast to Log_2(O/E) N/S ratios (Figure S5A; See Supplementary Note 3 for a detailed comparison
338 between methods).

339 To identify the contribution of the various types of non-synonymous variants to the negative
340 selection signal observed in the ENU-treated group, we stratified the non-synonymous variants
341 into stop retained, missense, start lost, stop gained, and stop lost variants and calculated their
342 respective Log_2(O/E) ratios (See Supplementary Note 4 for details on stop lost variants). This
343 revealed that missense variants made up the majority of non-synonymous variants in the ENU-
344 treated group (Figure 5E) and also that the negative selection signal was driven almost exclusively
345 by these variants (Figures 5F and S4H-J).

346 Next, we tested if variants in non-coding regions were under selection pressure. For this, we again
347 used the Variant Effect Predictor program to annotate SNVs in non-coding regions, i.e., introns,
348 intergenic, 5' untranslated regions (UTR), 3' UTR, and splice sites, and then quantified their
349 frequencies (Table S4, Methods)(*34*). As with the coding variants, non-coding variants showed
350 cycle-dependent increases in the ENU-treated group relative to the controls, with variants in
351 intronic and intergenic regions accounting for 49.9% and 35.1% of all non-coding variants,
352 respectively (Figure 5G). For ENU-treated groups at all cycles, intronic region variants Log_2(O/E)
353 ratios ≈ -0.08, significantly less than what would be expected by chance (Figure 5H, Table S5).
354 Since the vast majority of introns (with the exception of splice sites and other conserved non-
355 coding elements) do not contain functional sequences and are evolving at neutrality in the
356 germline, but are still subject to DNA repair such as TCR and MMR(*37, 38*), we interpret the
357 negative Log_2(O/E) ratios in these regions as a signal of DNA repair rather than negative selection.
358 In fact, as described at the beginning of this section, we use the intron Log_2(O/E) ratio to account
359 for the confounding effect of DNA repair on the Log_2(O/E) ratio. In intergenic regions for ENU-

360 treated groups at all cycles, we found that the Log_2(O/E) was ~0.13, significantly greater than what
361 would be expected by chance and indicative of mutation enrichment (Figure 5H, Table S5). Similar
362 to introns, the vast majority of intergenic regions lack functional sequences and are evolving at
363 neutrality in the germline. However, DNA repair activity in these regions is known to be
364 substantially reduced and likely explains the positive Log_2(O/E) ratios(*40, 41*). See also
365 Supplementary Note 5.

366 Out of all the non-coding variants, only Log_2(O/E) ratios of variants occurring in 3' UTR and
367 splice regions in the ENU-treated group at cycle 9 were significantly less than 0 (Figures 5H, S4K-
368 M, Table S5), which was indicative of negative selection and supported by previous studies on
369 germline selection pressures(*40, 42, 43*). The Log_2(O/E) ratios of 3' UTRs and splice sites in the
370 ENU-treated group at cycle 9 corresponded to ~20-25% reduction in mutation burden, a decrease
371 of 19.3±12.5 and 6.2±2.7 variants per cell, respectively (Figure 5I).

372

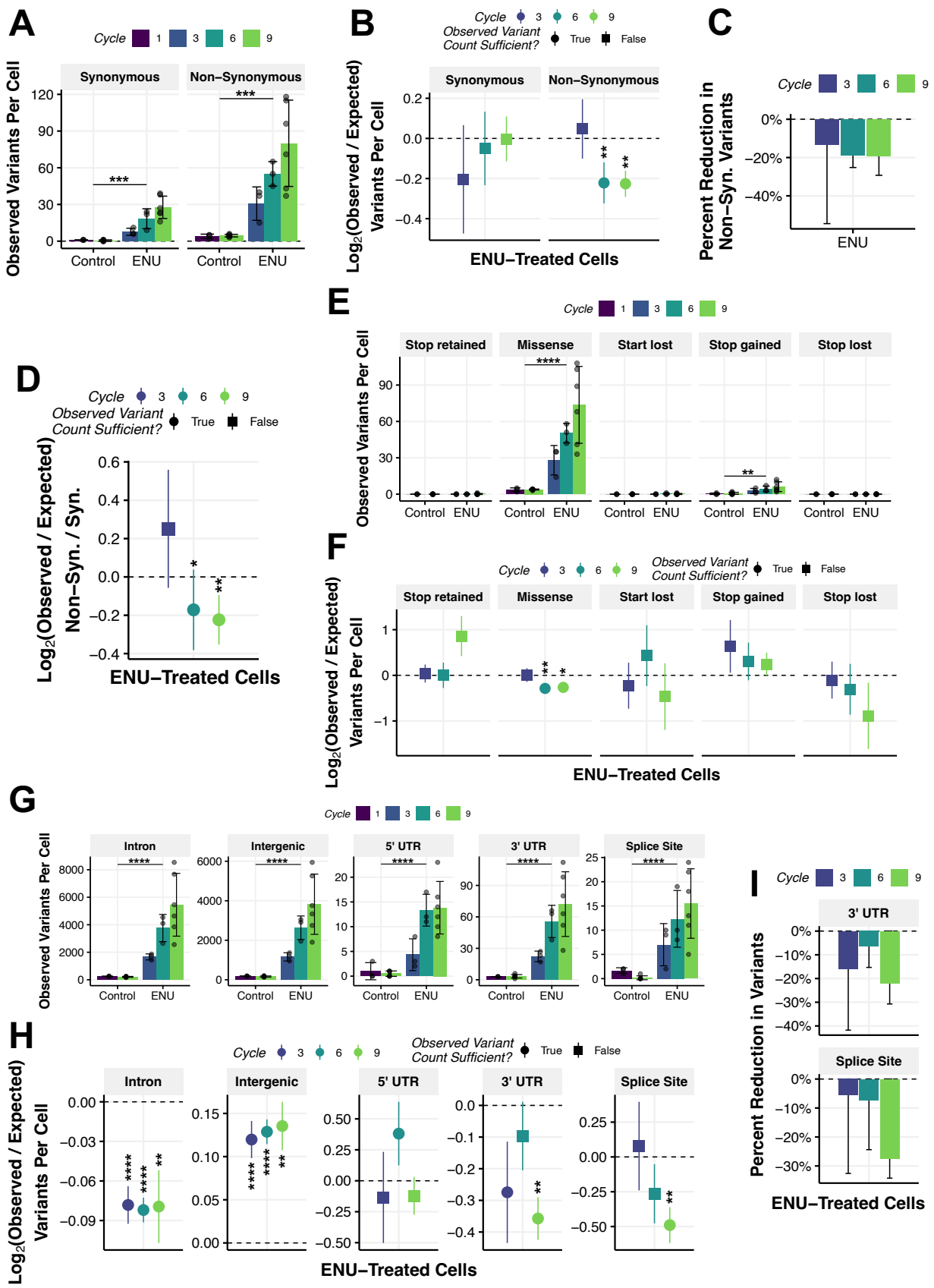

A
Cycle 1 3 6 9
Synonymous
Non–Synonymous
Observed Variants Per Cell
Control ENU
B
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU–Treated Cells
C
Cycle 3 6 9
Percent Reduction in Non–Syn. Variants
ENU
D
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Non–Syn. / Syn.
ENU–Treated Cells
E
Cycle 1 3 6 9
Stop retained
Missense
Start lost
Stop gained
Stop lost
Observed Variants Per Cell
Control ENU
F
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU–Treated Cells
G
Cycle 1 3 6 9
Intron
Intergenic
5' UTR
3' UTR
Splice Site
Observed Variants Per Cell
Control ENU
H
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU–Treated Cells
I
Cycle 3 6 9
3' UTR
Splice Site
Percent Reduction in Variants
ENU–Treated Cells

Figure 5. Negative Selection Against Damaging Coding and Non-Coding Variants in ENU-treated Cells.

(A) *Observed frequency of synonymous and non-synonymous variants per cell. See Table S3 for sample details and Table S5 for variant annotation. Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA and shown at very top of plot; p-value legend: ***: p <= 0.001.* ***(B)*** *Log_2(O/E) ratio of synonymous and non-synonymous variants for ENU-treated cells. See figures S4C, S4D, and S4E for results of without normalization, results of control cells, and results per cell, respectively. See Table S5 for detailed results. Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values to assess if the Log_2(O/E) ratio significantly deviates from 0 were estimated using a permutation test (n=10,000) but were only estimated if the frequency of observed variants allowed for sufficiency statistical power (Figure S4B; Table S5); p-value legend: **: p <= 0.01.* ***(C)*** *Estimated percent of non-synonymous variants out of total coding variants that are reduced in the ENU-treated group due to negative selection. Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; data represent mean ± S.D.* ***(D)*** *Log2O/E of the N/S ratio for ENU-treated cells. See figures S4I and S4J for results of control cells and per cell, respectively. Statistics: Same as in (B). p-value legend: *: p <= 0.05; **: p <= 0.01.* ***(E)*** *Observed frequency of various types of non-synonymous variants per cell. Statistics: Same as in (A). p-value legend: ****: p <= 0.0001.* ***(F)*** *Log2O/E of various types of non-synonymous variants for ENU-treated cells. See figures S4F, S4G, and S4H for results of without normalization, results of control cells, and results per cell, respectively. Statistics: Same as in (B). p-value legend: *: p <= 0.05; **: p <= 0.01.* ***(G)*** *Observed frequency of various types of non-coding variants per cell. Statistics: Same as in (A). ****: p <= 0.0001.* ***(H)*** *Log_2O/E of various types of non-coding variants for ENU-treated cells. See figures S5K, S5L, and S5M for results of control cells, without normalization, and per cell, respectively. See Supplementary Note 5 for discussion on results of the intergenic region. Statistics: Same as in (B). ****: p <= 0.0001; **: p <= 0.01.* ***(I)*** *Estimated percent reduction of variants in 3' UTR and splice region variants that are protected against in the ENU-treated group due to negative selection. Statistics: Same as in (C).*

Taken together, these results show that potentially damaging variants in both coding and a subset of non-coding regions did not accumulate to levels that would be expected in the context of the extremely high mutation burden induced by ENU treatment, and instead were subject to negative selection, resulting in lower mutation burdens within critical genes supporting essential cellular functions *in vitro*. This would explain how the population of cells exposed to repeated mutagen treatments was able to avoid a dramatic decline in growth rate, even as overall SNV burden continued to increase linearly throughout the experiment. Of note, the selection pressure analysis

410 was limited to mutations that were actually physically detected, which are ~20% of the ~56,000
411 estimated mutations at cycle 9 in the ENU-treated group (Figures 3A and S2H). Hence, these
412 results have relevance for the much smaller numbers of SNVs that are estimated to accumulate
413 during normal aging, e.g., from ~3,000 in human B lymphocytes to ~ 9,000 in human
414 hepatocytes(*6*).

415

416 Pathways Supporting Fibroblast Growth and Cell Identity are Protected from 417 Mutation Accumulation

418 To gain insight into selective pressures at broader levels associated with cellular function, we
419 investigated if gene sets active in proliferating human fibroblasts were under negative selection.
420 To do this, we analyzed the SNVs across annotated functional genomic regions relevant to human
421 lung fibroblasts (i.e. exons, upstream promoters, and enhancers) that were grouped into various
422 gene sets(*44-46*). We then calculated the frequency of SNVs within these regions as well as the
423 Log_2(O/E) ratios in a similar manner to the previous analyses (Methods).

424 First, we analyzed non-expressed genes and expressed genes as determined by our bulk mRNA-
425 sequencing data (Figures S1A-B; Table S7; Methods). In keeping with the previous analyses of
426 other regions (e.g. exonic and intronic regions), we found cycle-dependent increases of SNV
427 frequency in non-expressed and expressed regions for the ENU-treated group relative to the
428 controls (Figure 6A). Log_2(O/E) ratios of SNVs in expressed genes were significantly below 0 at
429 all cycles in the ENU-treated group, indicative of negative selection (Figures 6B and S6A-C, right
430 panels). In contrast, the Log_2(O/E) ratios of SNVs in non-expressed genes did not significantly
431 deviate from the expectation in the ENU-treated group at cycle 9, indicating neutral selection

432 (Figures 6B and S6A-C, left panels). Next, we further subdivided the expressed genes into
433 nonessential and essential gene sets based on previous *in vitro* screens (Table S7) and repeated the
434 same analysis as above (*47, 48*). This showed increased SNV frequency in the ENU-treated group
435 relative to the controls (Figure 6C), along with a Log_2(O/E) ratio significantly below 0 at cycle 9
436 in the ENU-treated group, indicative of negative selection (Figures 6D and S6D-F, right panels).
437 Of note, this is in agreement with global dN/dS ratios on truncating mutations across 17 genes
438 essential from a recent large-scale study sampling the normal oral epithelium(*49*).

439 As negative selection is expected to be generally weak due to the diploid nature of the human
440 genome(*39, 50*), but stronger in haploinsufficient regions (i.e. regions where both alleles are
441 required for function), we subdivided expressed genes into haplosufficient (pLI > 0.1) and
442 haploinsufficient (Table S7) (pLI < 0.9) gene sets based on a large scale study of protein-coding
443 genetic variation(*35*). As expected, analysis of these gene sets showed increased SNV frequency
444 in the ENU-treated group relative to the controls (Figure 6E). Log_2(O/E) ratios were significantly
445 below 0 in both haplosufficient and haploinsufficient gene sets for the ENU-treated cells at cycle
446 9. However, only the ENU-treated group at cycle 6 showed Log_2(O/E) ratios significantly below
447 0 for the haploinsufficient gene set and not the haplosufficient gene set, suggesting overall stronger
448 negative selection of haploinsufficient genes, as would be expected (Figures 6F and S6G-I). In
449 further support of this, when the Log_2(O/E) was compared between haplosufficient and
450 haploinsufficient gene sets for ENU-treated cells at cycle 9, we found that the Log_2(O/E) ratio was
451 considerably lower for the haploinsufficient genes (-0.37±0.16) as compared to the haplosufficient
452 genes (-0.28±0.04). In summary, these results suggest that damaging SNVs are selectively purged
453 from functional genomic regions based on expression status, essentiality, and dosage sensitivity,

454 extending our previous results on negative selection of damaging SNVs in coding and non-coding
455 regions.

456 Moving to the pathway level, we identified 399 gene ontology (GO) biological processes as
457 significantly enriched in the set of expressed genes from our bulk RNA-seq data (Figure S6J, Table
458 S8, Methods). As expected, the mean frequency of SNVs within each pathway showed significant
459 cycle-dependent increases of SNV frequency in the ENU-treated group relative to the controls
460 (Figure 6G). At the global level, Log_2(O/E) ratios of the identified pathways in the ENU-treated
461 group showed a cycle-dependent decrease, suggesting widespread negative selection at the
462 pathway level (Figures 6H, S6K-L, Table S9). At cycle 9 in the ENU-treated group, 6 pathways
463 showed Log_2(O/E) ratios significantly below 0 after multiple hypothesis correction, indicative of
464 negative selection. The pathways under negative selection were characterized as those related to
465 basic cellular processes supporting lung fibroblast cell growth *in vitro* and cellular identity, i.e.,
466 'response to oxidative stress' (GO:0006979), 'respiratory tube development' (GO:0030323),
467 'respiratory system development' (GO:0006979), 'regulation of response to DNA damage
468 stimulus' (GO:2001020), 'proteasome-mediated ubiquitin-dependent protein catabolic process'
469 (GO:0043161), and 'cellular response to chemical stress' (GO:0062197) (Figures 6I-K, Table S9).
470 To ensure that the pathways we had identified to be under negative selection were not false
471 positives, we repeated the pathway analysis, but employed bootstrapping to randomly shuffle
472 genes amongst pathways before calculating the Log_2(O/E) ratios (Figures S7A-C, Methods). The
473 results of this served as a negative control, which had no significant overlap with the pathways we
474 identified to be under significant negative slection pressure, providing a validation of our results
475 (Figures S7D-E).

476

477

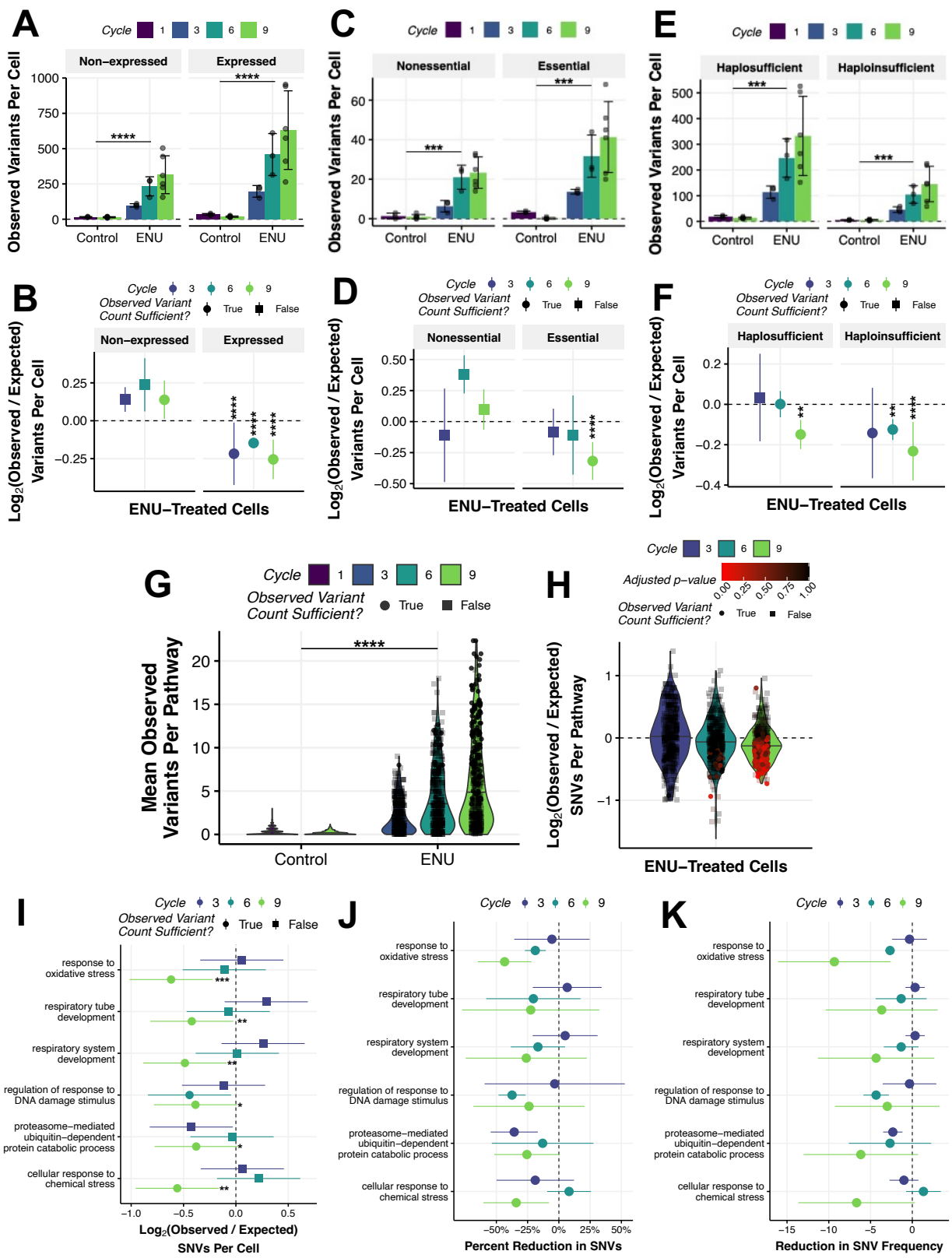

A
Cycle 1 3 6 9
Non–expressed
Expressed
Observed Variants Per Cell
Control
ENU
B
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU–Treated Cells
C
Nonessential
Essential
D
E
Haplosufficient
Haploinsufficient
F
G
Mean Observed Variants Per Pathway
H
Adjusted p–value
Log2(Observed / Expected) SNVs Per Pathway
I
response to oxidative stress
respiratory tube development
respiratory system development
regulation of response to DNA damage stimulus
proteasome–mediated ubiquitin–dependent protein catabolic process
cellular response to chemical stress
Log2(Observed / Expected) SNVs Per Cell
J
Percent Reduction in SNVs
K
Reduction in SNV Frequency

Figure 6. Pathways Supporting Fibroblast Growth and Cell Identity are Protected from Mutation Accumulation

(A) *Observed frequency of SNVs per cell in functional regions associated with genes that were determined to be non-expressed or expressed from bulk mRNA transcriptome data (Table S7). Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA; p-value legend: ****: p <= 0.0001.* ***(B)*** *Log_2(O/E) ratio of SNVs in functional regions associated with genes determined to be non-expressed or expressed expressed in bulk mRNA transcriptome data (Table S7). See figures S6A, S6B, and S6C for results without normalization, results of control cells, and results per cell, respectively. Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values to assess if the Log_2(O/E) ratio significantly deviates from 0 were estimated using a permutation test (n=10,000) but were only estimated if the frequency of observed variants allowed for sufficiency statistical power (Figure S4B; Table S5); p-value legend: ****: p <= 0.0001.* ***(C)*** *Observed frequency of SNVs per cell in functional regions associated with nonessential or essential gene sets expressed in bulk mRNA transcriptome data (Table S7). Statistics: Same in in (A); p-value legend: ***: p <= 0.001.* ***(D)*** *Log_2(O/E) ratio of SNVs in functional regions associated with nonessential or essential gene sets expressed in bulk mRNA transcriptome data (Table S7). See figures S6D, S7E, and S7F for results without normalization, results of control cells, and results per cell, respectively. Statistics: Same as in (B); p-value legend: ****: p <= 0.0001.* ***(E)*** *Observed frequency of SNVs per cell in functional regions associated with haplosufficient or haploinsufficient gene sets expressed in bulk mRNA transcriptome data (Table S7). Statistics: Same in in (A); p-value legend: ***: p <= 0.001.* ***(F)*** *Log_2(O/E) ratio of SNVs in functional regions associated with haplosufficient or haploinsufficient gene sets expressed in bulk mRNA transcriptome data (Table S7). See figures S6G, S6H, and S6I results without normalization, results of control cells, and results per cell, respectively. Statistics: Same as in (B); p-value legend: **: p <= 0.01; ****: p <= 0.0001.* ***(G)*** *Mean observed frequency of SNVs within the functional regions of pathways expressed in bulk mRNA transcriptome data (Table S8). See figure S6J for top 20 most enriched pathways. Statistics: n=399 pathways; P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA; p-value legend: ****: p <= 0.0001.* ***(H)*** *Mean Log_2(O/E) ratio of SNVs within the functional regions of pathways expressed in bulk mRNA transcriptome data (Table S8). See figure S6K and S6L for results of control cell and results without normalization, respectively. See table S9 for detailed results. Statistics: n=399 pathways; Otherwise, same as in (B), with the addition of Benjamini-Hochberg correction for multiple hypothesis testing of pathways.* ***(I-K)*** *Mean Log_2(O/E) ratio (I), estimated percent reduction in SNVs (J), and estimated reduction in SNV frequency (K) within functional regions of pathways found to be under significant negative selection (p <= 0.05) at cycle 9 in the ENU-treated group. Statistics: Same as in (B); ; p-value legend: *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001.*

In summary, these results suggest that in the context of very high mutation burden, pathways supporting oxidative stress, cellular identity, and protein homeostasis in proliferating cells *in vitro*, are protected from the accumulation of SNVs by negative selection. Importantly, these pathways

520 are biologically significant, as they appear to be relevant to the conditions of *in vitro* cell culture
521 and to the unique charecteristics of lung fibroblasts, highlighting the functional importance of
522 preventing mutation accumulation. These analyses show how negative selection allows a
523 population of dividing cells to survive supraphysiological mutation burden and underscores the
524 impact of random somatic SNVs on essential cellular functions.

525

526 # Discussion

527 Somatic mutations are known since the 1970s to be the main cause of cancer(*51*). Mainly because
528 of the exponential increase in cancer risk with age(*52*), aging itself has been suggested to be caused
529 by the accumulation of somatic mutations(*3, 8, 53*). Testing this hypothesis has long been
530 constrained by a lack of methods to quantitatively analyze somatic mutations in normal tissues
531 during aging. However, both single-cell and single-molecule methods have recently been
532 developed to accurately quantify randomly accumulating SNVs and INDELs. Using these
533 methods, we and others have demonstrated that somatic mutation burden is far higher than
534 expected based on the known germline mutation frequency and increases with age up to several
535 thousands of mutations per cell in aged subjects depending on the tissue(*6, 9*). However, a causal
536 relationship of increased mutation burden and age-related functional decline and disease risk
537 remains untested. Instead, attention has been focused on clonally amplified somatic mutations and
538 their causal relationship to non-cancerous, age-related diseases(*15, 54*).

539 In the present study, we tested for a possible limit to somatic mutation accumulation in a cell
540 culture model, in which high levels of somatic mutations were induced by multiple low doses of
541 the powerful mutagen ENU (Figure 1A). In this system, we were able to linearly increase the

542 somatic mutation burden to ~56,000 SNVs per cell, over 30-fold higher than untreated control
543 cells (Figure 3A). Remarkably, this supraphysiological mutation load only slightly impacted cell
544 growth and death rates (Figures 1B and 1E-H), suggesting that most cells can sustain an astounding
545 level of genetic information loss. The most straightforward explanation for these findings is that
546 the mammalian genome is highly robust, and that even very high numbers of point mutations are
547 unlikely to adversely affect cellular function. Indeed, this would be supported by results from a
548 recent study in which as many as ~30,000 SNVs were observed in intestinal crypts of elderly
549 patients with genetic defects in DNA mismatch repair(*55*).

550 An alternative explanation for the high tolerance of actively proliferating cells to mutation
551 accumulation is negative selection against mutations in genomic regions important for cell
552 survival. Evidence for negative selection against somatic mutations has been found in tumors(*56-*
553 *58*). However, in cancer, positive selection is the dominant force, arising from mutations in cancer
554 driver genes(*39*). Evidence for negative selection of somatic mutations in normal cells is scarce
555 because of the small numbers of mutations available as compared to tumors, which limits robust
556 statistical testing(*6*). Our present analysis of untreated control cells indeed suffered from this same
557 limitation, and we were unable to convincingly demonstrate negative selection of mutations in
558 these cells (Figure S4D, Table S5). However, our ENU-treated cells did yield a sufficiently high
559 number of SNVs to test for negative selection (Figure S2H). With this, we found multiple lines of
560 evidence for negative selection acting on potentially deleterious coding and non-coding variants.
561 While the frequency of both synonymous and non-synonymous variants increased linearly with
562 ENU treatment (Figure 5A), deeper analysis using observed vs. expected ($Log_2(O/E)$) ratios
563 revealed a significantly lower frequency of high impact, non-synonymous variants, especially
564 missense variants, in the ENU-treated cells as compared to our null model (Figures 5B-C). This

565 suggests that an actively proliferating population of fibroblasts has the capacity to selectively purge
566 the most harmful coding mutations from their genomes, preserving cellular fitness despite the
567 extremely high number of random mutations accumulating genome wide, and agrees with the
568 selection pressures found from analyzing germline mutations(*35, 59-61*). This main finding was
569 recently validated in an *Msh2*$^{-/-}$ mouse model, which found negative selection against somatic stop
570 lost variants(*26*). Additionaly, negative selection was also detected in non-coding regions for
571 variants occurring in 3' UTR and splice sites (which collectively made up ~1.1% of all the
572 observed non-coding variants) (Figures 5G-I), which is in agreement with recent literature
573 analyzing selective pressures in these regions (*43*).

574 Our analysis of gene sets and pathways revealed negative selection to also be acting at a broader
575 level, highlighting the biological significance. In the primary lung fibroblasts treated with ENU,
576 we found evidence of negative selection in functional genomic regions of expressed, essential and
577 haploinsufficient genes (Figures 6A-F). This finding was recently recapitulated in a population
578 scale study on somatic mutations in oral epithelium cells that observed negative selection on
579 essential genes, which overlaps with the sets of essential and haploinsufficient genes that was used
580 in our study (Figures 6C-F)(*62*). Moreover, signals of negative slection were identified in actively
581 expressed pathways, such as oxidative stress, respiratory system development, and response to
582 DNA damage (Figures 6I-K), which were biologically relevant to our *in vitro* experimental
583 conditions and the lung fibroblast cell type that was used. For example, selection was observed
584 against genes involved in respiratory system development, which include genes such *FGF10,*
585 *FOXF1,* and *COL3A1* that are known to be crucial for lung fibroblast identity and function(*63-*
586 *65*). This suggests that despite the massive somatic mutational burden across the genome overall,

587 random somatic mutations that occur in pathways critical for cellular survival do indeed have a
588 damaging effect.

589 Our work provides conclusive evidence that very high numbers of somatic mutations, collectively,
590 may have a functional impact on normal cells *independent* of clonally amplified mutations.
591 However, because a supraphysiological amount of mutations was required to detect negative
592 selection, this puts into question whether there is an effect of negative selection on the much lower
593 numbers of somatic mutations that accumulate during aging. In this respect, it should be noted that
594 the actual number of SNVs physically observed wass much smaller that the ~56,000 estimated
595 based on genome coverage and sensitivity, i.e., no more than ~12,000 (Figures 3A and S2H). This
596 is not that much above the ~2,000 and ~5,000 SNVs per cell, depending on the cell type analyzed,
597 that have been estimated to have accumulated at old age(*6*). Moreover, we did not observe an
598 association with the strength of negative selection and mutation burden (Figure 5B). Hence, it is
599 conceivable that negative selection is a general mechanism to eliminate somatic mutations that
600 adversely impact cell function, similar to purifying selection in the germline.

601 Another important aspect when considering the implications of negative selection are the
602 differences between our *in vitro* model and the *in vivo* context. While negative selection in our
603 model acted on cell proliferation, survival, and apoptosis, immune surveillance (i.e. natural killer
604 cells and macrophages) would provide additional targets for *in vivo* elimination of fibroblasts
605 displaying neoantigens or excreting senescence-associated factors. Moreover, in normal adult
606 subjects, cell division rates are very low with few exceptions, most notably the lymphoid and
607 intestinal systems(*19*). In proliferating cells, such as human B lymphocytes, we have previously
608 demonstrated that SNVs accumulate with age much slower in the functional part of the genome
609 than in the genome overall, indicating negative selection of mutations arising during the process

610 of B cell aging(*66*). By contrast, somatic mutations accumulating in non-dividing adult cells, such
611 as neurons or cardiomyocytes, cannot be selected against through proliferation or apoptosis, which
612 is evidenced by excess accumulation of damaging mutations(*12, 67*). This suggest that age-related
613 somatic mutation accumulation would have the greatest impact on non-proliferative cells, which
614 may be a determinant in species lifespan as predicted by the disposable soma theory(*68*).

615 As a limitation of our study, we should point out that we only analyzed SNVs, and the few small
616 insertions and deletions induced by ENU. However, SNVs are much less likely to have a functional
617 impact than genome structural variants (SVs), including large deletions, translocations, and copy
618 number variants. Indeed, this is clearly indicated for INDELs, which in our present study did not
619 linearly accumulate during the ENU treatment (Figures S2I-J). In support of this, we recently
620 conducted a study where the limit to the number of INDELs, ~16,000 per cell, was determined by
621 profiling somatic mutation accumulation in fibroblasts of DNA mismatch repair deficient *Msh2*$^{-/-}$
622 (*26*). Thus far, there is very little information about somatic SVs, which cannot be detected
623 accurately, either using single-cell or single-molecule methods(*69*). While their relative frequency
624 is much lower than that of SNVs, SVs can have a major phenotypic impact, as indicated by
625 germline SVs in human diseases(*70*). SVs are much more likely to adversely affect cell function
626 than SNVs, but unless they result in cell death, a mechanism to select against them in non-dividing
627 cells *in vivo* during aging is difficult to imagine.

628 Future efforts should extend the framework presented in this study to additional cell types,
629 especially quiescent populations, and incorporate methods that capture structural variation.
630 Integrating mutational landscapes with functional readouts *in vivo* will be crucial for determining
631 how the balance between information loss and negative selection shapes organismal aging and
632 disease.

633 # Methods

634 ## Cell culture and treatment

635 Human IMR-90 primary fibroblasts were obtained from the American Tissue Culture Collection
636 (ATCC, Lot 70020198). They were cultured in DMEM (Dulbecco's Modified Eagle Medium,
637 from ATCC) supplemented with 10% fetal bovine serum (FBS, Gibco/Life Technologies) and 1%
638 Penicillin-Streptomycin solution (Gibco/Life Technologies) at 3% oxygen and 37°C. Cells
639 obtained from ATCC were at passage 11 at arrival and were cultured for an additional 5 passages
640 before the start of the experiment. At the beginning of each cycle, $1*10^6$ human primary lung
641 fibroblasts (IMR-90) cells were counted using a Countess 3 (Fisher Scientific), plated, and allowed
642 to expand for 24 hours. Cells were then treated with 50 µg/mL N-ethyl-N-nitrosourea (ENU) and
643 allowed to recover for 3 days. Cells were then collected and assayed for cell number and viability
644 for the 3-day time point. Of the collected cells, $1*10^6$ cells were re-plated and then allowed another
645 4 days to recover and expand until they were collected to assess cell number and viability for the
646 7-day time point. This procedure was repeated for 9 consecutive cycles over the course of ~10
647 weeks. This experiment was repeated 3 separate times to create 3 independent biological replicates.
648 Cells were isolated at the 7-day time point at cycles 1, 3, 6, and 9 for somatic mutation analysis
649 and at cycles 1 and 9 for transcriptome analysis.

650

651 ## Cell growth and apoptosis assays

652 Cells were counted using the ViaCount assay (Cytek) on a Guava EasyCyte flow cytometer to
653 assess cell number and viability at each time point. Apoptosis rates were determined utilizing the

654 Nexin assay (Cytek) on the same machine. Population doublings (PDs) is defined as how many
655 times a population of cells has doubled over a given period of time, was was calculated as:

656 $$PD = 3.32 * (log10\,(number\ of\ cells\ counted) - log10\,(number\ of\ seeded\ cells)).$$

657

658 **Bulk mRNA transcriptome analysis**

659 RNA was extracted from frozen cell pellets using the Zymo DNA/RNA extraction kit (Microprep
660 plus D7005) according to the manufacturer's instructions and quantified using a Qubit kit (Thermo
661 Fisher Scientific). Libraries for mRNA analysis were prepared and sequenced using 150 paired-
662 end mode on an Illumina Novaseq platform by Novogene Corp Inc., CA. Raw sequence reads
663 were processed using the nf-core RNA-Seq pipeline (version 3.1.0) to obtain gene expression
664 levels(*71, 72*). Gene filtering was performed using the zPKM program (Figure S1A)(*73*). Count
665 normalization and differential expression testing was performed using DESeq2 where the design
666 was set to test for the effect of the cycle-condition variable while controlling for the batch variable
667 (Supplementary Note 1)(*74*). Gene set enrichment analysis (GSEA) was performed using the
668 clusterProfiler program (*75*) along with KEGG DNA repair, KEGG apoptosis (*76*) and the
669 SenMayo gene sets(*77*).

670

671 **Single-cell isolation, whole-genome amplification, library preparation, and**
672 **sequencing**

673 Single-cells were isolated using the CellRaft AIR system (Cell Microsystems). The CellRaft array
674 was prepared following the manufacturer's instruction and washed three times for 3 minutes with

675 warm PBS before addition of cell suspension. 1,000 cells were seeded in 1 mL for one CellRaft
676 array. Cells were allowed to settle for 3-4 hours at 37°C before isolation. Then, the raft was
677 carefully washed with warm medium to remove potential debris and dead cells. Individual rafts
678 containing one cell were deposited in PCR tubes with 2.5 µL of PBS and stored at -80°C until
679 further processing. Single-cell whole-genome amplification and library preparation for WGS was
680 performed as previously described [12]. In brief, single-cell whole genomes were amplified using
681 SCMDA and subject to quality control using a locus dropout test(*78*). Libraries were prepared
682 using a NEBNext Ultra II FS kit (NEB). Library quantity and size were assessed with a Qubit kit
683 and TapeStation, respectively. Libraries were sequenced using the 150 paired-end mode on an
684 Illumina Novoseq platform by Novogene Corp Inc., CA.

685

686 **Single-cell whole-genome read alignment and mutation calling**

687 WGS data processing and quality control was performed using as previously described(*24*).
688 Briefly, raw sequencing data underwent quality control using FastQC (*79*) and Trim Galore(*80*),
689 followed by alignment to the human hg19 reference genome using the GATK best practices
690 pipeline(*81*). This included marking adapters, sequence alignment, base quality score
691 recalibration, and variant calling for germline variants using GATK HaplotypeCaller. SCcaller
692 (version 2.0) (*24*)was then employed to call somatic single-nucleotide variations (SNVs) and small
693 insertions and deletions (INDELs), adjusting for allelic amplification bias using known germline
694 variants. Estimated mutation burden was calculated as follows:

695 $$\text{Estimated somatic SNVs|INDELs per cell} = \frac{\text{observed somatic SNVs|INDELs}}{\frac{surveyed\ genome\ (bp) * sensitivity}{size\ of\ genome\ (bp)}}$$

696 **Mutation filtering and identification of technical artifacts in single-cell whole-**
697 **genome mutation data**

698 Overlapping SNVs and INDELs were observed amongst the single-cells from batches 2 and 3
699 (Figures S2D-E; Table S3). Since this was not observed in the first batch, it is unlikely that these
700 mutations are the result of clonal amplification. Mutations that were shared between all cells within
701 a batch were thus considered to be germline and were filtered out prior to downstream analysis
702 (Figures S2F-G).

703 Technical artifacts inherent to single-cell MDA library preparation steps, such as overheating
704 during cell lysis or whole genome amplification, can be identified by their specific mutational
705 signatures scE and scF(*82*). We checked for the presence of these signatures and found scF to be
706 present in the signature extracted from control cells at cycle 1 (Figure S3A-C). The scF technical
707 artifact is characterized by higher than expected cytosine-to-thymine (C>T) transversions at G<u>C</u>N
708 contexts (Figures SF-G). Thus, control cells from cycle 1 were excluded from the downstream
709 mutational signature analyses.

710 In an attempt to account for the C>T transition artifacts in control cells at cycle 1, we compared
711 the proportion of the C>T transitions in control cells at cycle 1 to control cells at cycle 9, which
712 do not show any mutational signatures with similarity to the scF sequencing artifact mutational
713 signature (*82*) and have mutation spectra previously observed for IMR90 cells(*83*). We found that
714 this artifact increased the proportion of C>T transitions by 18%±1%. With this, we estimated the
715 amount of artifactual C>T transitions in control cells at cycle 1, subtracted this from the observed
716 mutation burden, and then re-estimated the SNVs per cell (Figure S3D, Table S3).

717

718 ## Mutation signature analysis

719 The MutationalPatterns program (*84*) was used to perform mutational spectra and signature
720 analysis. We obtained a diverse set of samples from SomaMutDB (*85*) to use as the background
721 reference mutational signature analysis. These were single-cell whole genome sequencing samples
722 from 10 fibroblast cells(*23*), 199 brain cells(*30*),and 53 lung cells(*29*), where genomic DNA was
723 amplified with either Multiple Displacement Amplification (MDA) or Primary Template-directed
724 Amplification (PTA) (Figure S3E). To determine the number of mutational signatures to be
725 extracted from the data, a non-negative matrix factorization (NMF) rank survey was employed
726 which resulted in 2 factors (Figures S3F). *De novo* signatures were extracted and then compared
727 to those from the COSMIC and SIGNAL databases using cosine similarity(*86, 87*).

728

729 ## Variant annotation

730 Observed and expected mutations from each single-cell (or simukation instance of a single-cell in
731 the case of expected mutations) were annotated using the Ensembl Variant Effect Prediction
732 program (version 110.1) and the given 'Consequences' annotation was used for downstream
733 analyses(*34*). In the case where mutations were predicted to lead to more than 1 unique
734 consequence (i.e. multiple entries of the same mutation), the highest ranking annotation that
735 occurred in the resulting output was kept which resulted in a 1:1 mapping of mutations and
736 consequences. In the case where mutations occurred in a region that overlapped multiple possible
737 consequences (e.g. splice region variant and intron variant), we classified this mutation to be
738 consistent with the consequence that occurred first in this annotation (e.g. splice region, if multiple
739 annotation was splice region variant and intron variant). For mutations annotated as splice variants

740 (splice_acceptor_variant, splice_donor_variant, splice_donor_5th_base_variant,
741 splice_region_variant,splice_donor_region_variant, splice_polypyrimidine_tract_variant), we
742 classified these mutations as 'splice.' Synonymous variants were classified as mutations annotated
743 as synonymous_variant or stop_retained_variant. Non-synonymous variants were classified as
744 mutations annotated as missense_variant, start_lost, stop_gained, or stop_lost.

745

746 **Selection analysis of annotated variants**

747 Selection analysis was performed by first simulating 10,000 instances of each single-cell in our
748 dataset to be used as the null expectation (i.e. neutral selection) where mutations are randomly re-
749 distributed throughout the genome. This was done using the SigProfilerSimulator program(*33*),
750 where for each single-cell, observed mutations were randomly re-distributed across the genome
751 while accounting for cell-specific sequencing coverage, mutational signatures, and nucleotide
752 contexts.

753 Log_2(observed/expected) ratios (written hereafter as Log_2(O/E)) were calculated for each single-
754 cell or by group (mutations from all cells within a group are pooled), by first summing the
755 frequency of annotated consequences for observed and expected mutations. The Log_2(O/E) ratio
756 was then calculated with respect to each simulation instance (resulting in 10,000 Log_2(O/E) ratios
757 per cell or group) using the following equation:

758 $$Log_2(O/E) = Log_2 \frac{(observed\ consequence\ freq. + 1)}{(simulated\ consequence\ freq. + 1)}$$

759

760 A permutation test was then used to determine the p-value of the observed consequence frequency
761 where the simulated consequence frequency served as the null distribution and is as follows:

762 $$p = \frac{1}{n}\sum_{i=1}^{n} I\,(simulated\ consequence\ freq._i \geq observed\ consequence\ freq.)$$

763 Where p is the permutation test p-value, n is the number of permutations (i.e. 10,000), i is the i-th
764 permutation, and I is the indicator function, which equals 1 if the condition inside is true, and 0
765 otherwise.

766 We performed a power analysis to determine how many observed variants are required to obtain
767 reliable statistical results at varying Log_2(O/E) ratios. To do this, we used the mutation data from
768 ENU-treated cells to obtain the relationship between the standard deviation in the number of
769 expected variants given the number of observed variants for common types of variants that were
770 identified. We fit a gamma GLM with a log link to model standard deviation as a function of log-
771 transformed observed variant count. The model showed excellent fit (residual deviance = 3.19 on
772 144 df vs. null deviance = 147.37 on 145 df; dispersion = 0.020), with log-transformed observed
773 variant count as a highly significant predictor irrespective of the variant type ($p < 2 \times 10^{-16}$). With
774 this, we then calculated Log_2(O/E) ratios by simulating 10,000 instances of expected variant counts
775 while varying the observed variant counts from 1-1,000. We then varied the Log_2(O/E) ratios from
776 0-1 by multiplying the observed variant count by a scaling factor before calculating the Log_2(O/E)
777 ratio. Finally, we calculated p-values for the Log_2(O/E) ratios using a permutation test as described
778 previously.

779 The power analysis demonstrated how varying the observed variant count affects the p-value of
780 the Log_2(O/E) ratio at different levels. Importantly, this analysis provided us with estimates of how
781 many observed variants were required to achieve sufficient power (p-value<0.05) (i.e. obtain

782 reliable statistical results) given a specific Log_2(O/E) (Figure S4B, solid line). This threshold was
783 then applied to the Log_2(O/E) results, at both the levels of variant types and pathways, to determine
784 which of these were amenable to further statistical testing where the results of this can be seen in
785 the figure legends (e.g. Figure 5B; Table S5). For the Log_2(O/E) non-synonymous/synonymous
786 calculation, we determined if these were amenable to further statistical testing if either the non-
787 synonymous or synonymous Log_2(O/E) results in each group reached the established observed
788 variant threshold (Table S5).

789 When assessing for mutation selection, it is also necessary to account for the cell-intrinsic bias in
790 mutation frequency due to DNA repair mechanisms (i.e. transcription coupled repair), which is
791 known to reduce mutation frequency in transcribed regions and is agnostic to damage that would
792 result in synonymous or non-synonymous variants(*88*). This bias in mutation frequency was
793 indeed observed in our data and consistent with a depletion of mutations due to DNA repair (Figure
794 5H). Thus, we normalized the Log_2(O/E) ratio of coding variants (e.g. synonymous, missense, etc.)
795 and non-coding variants (5'/3' UTR and splice region) by the Log_2(O/E) ratio of mutations which
796 occurred within intronic regions and not in splice regions. We chose to use this normalization
797 because the occurrence of both exonic and intronic mutations is dependent on transcription coupled
798 repair (*37*) however, since intronic mutations do not alter the translated protein (with the exception
799 of splice site variants), they are not subject to selective mechanisms. In addition, the majority of
800 variants in our data occur within introns (~46.7%, Figure 5G), allowing us to make more confident
801 Log_2(O/E) ratio for these mutations as compared to using the Log_2(O/E) ratio of synonymous
802 mutations which only made up ~0.23% of all variants. A comparison of unnormalized and
803 normalized Log_2(O/E) ratio for non-synonymous and synonymous variants is shown for example

804 by comparing Figures 5B to S4C, where the effect of normalization can be seen to slightly increase
805 the Log_2(O/E) ratio.

806 To estimate the number of variants that are protected against resulting from negative selection, we
807 subtracted the average number of variants from simulated samples from the number of variants in
808 the respective observed sample.

809

810 dN/dS analysis

811 For an alternative measure of selection, we employed the dndscv R package(*39*), which utilizes a
812 trinucleotide context-dependent substitution model to estimate dN/dS ratios. The analysis was
813 performed using the hg19 reference genome. We used default settings with the exception of
814 disabling the filtering of hypermutated samples and genes (max_coding_muts_per_sample = Inf
815 and max_muts_per_gene_per_sample = Inf) to retain all mutation data in the cohort.

816

817 Selection analysis of gene sets and pathways

818 Non-expressed and expressed genes were determined using bulk mRNA transcriptome data from
819 cycle 9 control and ENU-treated groups using the zPKM program (Figure S2A; Table S7)(*73*).
820 Nonessential and essential control gene sets were obtained from the DepMap database
821 (depmap.org)(*89*), where the nonessential gene set was the negative controls from Hart *et al* (2015)
822 and the essential gene set was the intersection of the the essential gene sets resulting from the Hart
823 *et al* (2015) and Blomen *et al* (2014) studies(*48, 90*). These gene sets were additionally filtered for
824 expressed genes by overlapping them with the list of expressed genes we determined previously

825 (Table S7). Haplosufficient and haploinsufficient gene sets were obtained from the ExAC database
826 where haplosufficient genes were defined as pLI < 0.1 and haploinsufficient genes were defined
827 as pLI > 0.9(*35*). These gene sets were also filtered for expressed genes in the same manner as
828 essential genes (Table S7). Pathway level gene sets were obtained by performing over
829 representation analysis for gene ontology biological processes using the clusterProfiler program
830 (*75*) on the list of expressed genes that determined previously (Table S8). For all gene sets,
831 genomic coordinates of exons, introns, promoters (5KB upstream of transcription start site), and
832 annotated enhancers (*44*) were retrieved using the GenomicRanges (version 1.60.0) software(*91*).

833 For each gene set, two Log_2(O/E) ratios were calculated. One for the version of the gene set where
834 the intronic regions were included, and the other for the version of the gene set intronic regions
835 were excluded. To account for DNA repair (as previously described for Log_2(O/E) calculation of
836 coding and non-coding variants), the Log_2(O/E) ratios of each gene set was normalized by
837 subtracting Log_2(O/E) ratio of the version that included introns from the version that excluded
838 introns. A permutation test was then used to estimate p-values, as described above, and were
839 adjusted for multiple hypothesis testing using the Benjamini-Hochberg method.

840 To serve as a negative control for the pathway analysis, for each enriched biological process gene
841 set, we randomly shuffled the genes between all other identified biological processes gene sets and
842 created 10,000 randomly permuted gene sets. Average Log_2(O/E) ratios were then calculated for
843 the 10,000 permutations of each gene set followed by p-value estimation, as described above.

844

845

Data and code availability

Raw sequencing data has been deposited in the NCBI Sequence Read Archive under BioProject accession PRJNA1129131. Processed gene expression data has been deposited in the NCBI Gene Expression Omnibus database under accession GSE271867. Processed variant call files have been deposited in the NCBI dbSNP database and will be made available upon acceptance. Scripts used for processing and analyzing data have been deposited at:

figshare.com/articles/software/Negative_selection_allows_human_primary_fibroblasts_to_tolerate_high_somatic_mutation_loads_induced_by_N-ethyl-N-nitrosourea/26110453

Competing interests

M.L., A.Y.M., X.D., and J.V. are co-founders and shareholders of SingulOmics Corp. A.Y.M., and J.V. are co-founders of MutagenTech Inc. Other authors declare no conflict of interest.

Acknowledgements

We would like to acknowledge the members of the Vijg and Sidoli labs for helpful discussions during the preparation of the manuscript.

Author Contributions

A.Y.M., and J.V. conceived this study and designed the experiments. J.H., M.S., and A.Y.M performed the experiments. R.C., J.H., and S.S. analyzed the data. X.D. provided guidance on the data analysis. R.C., J.H., and J.V. wrote the manuscript.

Funding

This study was supported by National Institute of Health grants: T32GM007491 (R.C.) T32AG023475 (R.C.), U19AG056278 (J.V.), U01HL145560 (J.V.), U01ES029519 (J.V.), P01AG017242 (J.V.), P01AG047200 (J.V.), RF1AG068908 (J.V.), and P30AG038072 (J.V.). Other funding sources: US Department of Defense grant BC180689P1 (J.V.), Glenn Foundation for Medical Research Postdoctoral Fellowships in Aging Research (J.H.), The Michael J. Fox Foundation (J.V.), and The Paul F. Glenn Center for the Biology of Human Aging at the Albert Einstein College of Medicine (J.V.).

References

1. C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, G. Kroemer, Hallmarks of aging: An expanding universe. *Cell* **186**, 243-278 (2023).
2. B. Schumacher, J. Pothof, J. Vijg, J. H. J. Hoeijmakers, The central role of DNA damage in the ageing process. *Nature* **592**, 695-703 (2021).
3. G. Failla, The aging process and cancerogenesis. *Ann N Y Acad Sci* **71**, 1124-1140 (1958).
4. L. Szilard, ON THE NATURE OF THE AGING PROCESS. *Proc Natl Acad Sci U S A* **45**, 30-45 (1959).
5. E. Laconi, F. Marongiu, J. DeGregori, Cancer as a disease of old age: changing mutational and microenvironmental landscapes. *Br J Cancer* **122**, 943-952 (2020).
6. P. Ren, X. Dong, J. Vijg, Age-related somatic mutation burden in human tissues. *Front Aging* **3**, 1018119 (2022).

892 7. I. Franco, M. Eriksson, Reverting to old theories of ageing with new evidence for the role
893 of somatic mutations. *Nat Rev Genet* **23**, 645-646 (2022).
894 8. I. Franco, G. Revêchon, M. Eriksson, Challenges of proving a causal role of somatic
895 mutations in the aging process. *Aging Cell* **21**, e13613 (2022).
896 9. I. Franco *et al.*, Whole genome DNA sequencing provides an atlas of somatic mutagenesis
897 in healthy human cells and identifies a tumor-prone cell type. *Genome Biol* **20**, 285 (2019).
898 10. I. Franco *et al.*, Somatic mutagenesis in satellite cells associates with human skeletal
899 muscle aging. *Nat Commun* **9**, 800 (2018).
900 11. M. A. Lodato *et al.*, Aging and neurodegeneration are associated with increased mutations
901 in single human neurons. *Science* **359**, 555-559 (2018).
902 12. S. Choudhury *et al.*, Somatic mutations in single human cardiomyocytes reveal age-
903 associated DNA damage and widespread oxidative genotoxicity. *Nature Aging* **2**, 714-725
904 (2022).
905 13. R. P. Erickson, Somatic gene mutation and human disease other than cancer: an update.
906 *Mutat Res* **705**, 96-106 (2010).
907 14. C. Li, S. M. Williams, Human Somatic Variation: It's Not Just for Cancer Anymore.
908 *Current Genetic Medicine Reports* **1**, 212-218 (2013).
909 15. A. Abyzov *et al.*, One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic
910 mutational load with patterns that suggest proliferative origin. *Genome Res* **27**, 512-523
911 (2017).
912 16. C. C. Laurie *et al.*, Detectable clonal mosaicism from birth to old age and its relationship
913 to cancer. *Nat Genet* **44**, 642-650 (2012).
914 17. A. Abyzov *et al.*, Somatic copy number mosaicism in human skin revealed by induced
915 pluripotent stem cells. *Nature* **492**, 438-442 (2012).
916 18. S. Jaiswal *et al.*, Age-related clonal hematopoiesis associated with adverse outcomes. *N*
917 *Engl J Med* **371**, 2488-2498 (2014).
918 19. R. Sender, R. Milo, The distribution of cellular turnover in the human body. *Nat Med* **27**,
919 45-48 (2021).
920 20. D. H. Lackner, M. T. Hayashi, A. J. Cesare, J. Karlseder, A genomics approach identifies
921 senescence-specific gene expression regulation. *Aging Cell* **13**, 946-950 (2014).
922 21. D. Saul *et al.*, A new gene set identifies senescent cells and predicts senescence-associated
923 pathways across tissues. *Nature Communications* **13**, 4827 (2022).
924 22. I. Barbaric, S. Wells, A. Russ, T. N. Dear, Spectrum of ENU-induced mutations in
925 phenotype-driven and gene-driven screens in the mouse. *Environ Mol Mutagen* **48**, 124-
926 142 (2007).
927 23. X. Dong *et al.*, Accurate identification of single-nucleotide variants in whole-genome-
928 amplified single cells. *Nat Methods* **14**, 491-493 (2017).
929 24. L. Zhang *et al.*, Analyzing somatic mutations by single-cell whole-genome sequencing.
930 *Nat Protoc* **19**, 487-516 (2024).
931 25. D. E. Watson, M. L. Cunningham, K. R. Tindall, Spontaneous and ENU-induced mutation
932 spectra at the cII locus in Big Blue Rat2 embryonic fibroblasts. *Mutagenesis* **13**, 487-497
933 (1998).
934 26. L. Zhang *et al.*, Divergent accumulation patterns of SNVs and INDELs reveal negative
935 selection in noncancerous cells. *Innovation (Camb)* **6**, 101008 (2025).
936 27. A. Y. Maslov *et al.*, Single-molecule, quantitative detection of low-abundance somatic
937 mutations by high-throughput sequencing. *Sci Adv* **8**, eabm3259 (2022).

938 28. L. B. Alexandrov *et al.*, The repertoire of mutational signatures in human cancer. *Nature*
939 **578**, 94-101 (2020).
940 29. Z. Huang *et al.*, Single-cell analysis of somatic mutations in human bronchial epithelial
941 cells in relation to aging and smoking. *Nat Genet* **54**, 492-498 (2022).
942 30. M. B. Miller *et al.*, Somatic genomic changes in single Alzheimer's disease neurons. *Nature*
943 **604**, 714-722 (2022).
944 31. J. E. Kucab *et al.*, A Compendium of Mutational Signatures of Environmental Agents. *Cell*
945 **177**, 821-836 e816 (2019).
946 32. T. Hwang *et al.*, Comprehensive whole-genome sequencing reveals origins of mutational
947 signatures associated with aging, mismatch repair deficiency and temozolomide
948 chemotherapy. *Nucleic Acids Res* **53**, (2025).
949 33. E. N. Bergstrom, M. Barnes, I. Martincorena, L. B. Alexandrov, Generating realistic null
950 hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC*
951 *Bioinformatics* **21**, 438 (2020).
952 34. W. McLaren *et al.*, The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
953 35. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**,
954 285-291 (2016).
955 36. C. Greenman, R. Wooster, P. A. Futreal, M. R. Stratton, D. F. Easton, Statistical analysis
956 of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187-2198 (2006).
957 37. E. E. Heilbrun, M. Merav, S. Adar, Exons and introns exhibit transcriptional strand
958 asymmetry of dinucleotide distribution, damage formation and DNA repair. *NAR Genom*
959 *Bioinform* **3**, lqab020 (2021).
960 38. J. Frigola *et al.*, Reduced mutation rate in exons due to differential mismatch repair. *Nat*
961 *Genet* **49**, 1684-1692 (2017).
962 39. I. Martincorena *et al.*, Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*
963 **171**, 1029-1041.e1021 (2017).
964 40. N. Dukler, M. R. Mughal, R. Ramani, Y. F. Huang, A. Siepel, Extreme purifying selection
965 against point mutations in the human genome. *Nat Commun* **13**, 4312 (2022).
966 41. A. A. Yurchenko *et al.*, Genomic mutation landscape of skin cancers from DNA repair-
967 deficient xeroderma pigmentosum patients. *Nat Commun* **14**, 2561 (2023).
968 42. S. V. Denisov *et al.*, Weak negative and positive selection and the drift load at splice sites.
969 *Genome Biol Evol* **6**, 1437-1447 (2014).
970 43. S. D. Findlay, L. Romo, C. B. Burge, Quantifying negative selection in human 3' UTRs
971 uncovers constrained targets of RNA-binding proteins. *Nat Commun* **15**, 85 (2024).
972 44. C. A. Boix, B. T. James, Y. P. Park, W. Meuleman, M. Kellis, Regulatory genomic
973 circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300-307 (2021).
974 45. M. B. Carlson M. (R package version 3.2.2. , 2015).
975 46. C. A. Boix. (2023).
976 47. T. Wang *et al.*, Identification and characterization of essential genes in the human genome.
977 *Science* **350**, 1096-1101 (2015).
978 48. T. Hart *et al.*, High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-
979 Specific Cancer Liabilities. *Cell* **163**, 1515-1526 (2015).
980 49. A. R. J. Lawson *et al.*, Somatic mutation and selection at epidemiological scale. *medRxiv*,
981 2024.2010.2030.24316422 (2024).
982 50. I. E. Vorontsov *et al.*, Negative selection maintains transcription factor binding motifs in
983 human cancer. *BMC Genomics* **17 Suppl 2**, 395 (2016).

984 51. P. C. Nowell, The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).
985 52. *US cancer statistics: 1999–2009 incidence and mortality web-based report.* (2013
986 www.cdc.gov/uscs).
987 53. F. M. Burnet, Intrinsic mutagenesis: a genetic basis of ageing. *Pathology* **6**, 1-11 (1974).
988 54. S. W. K. Ng *et al.*, Convergent somatic mutations in metabolism genes in chronic liver
989 disease. *Nature* **598**, 473-478 (2021).
990 55. P. S. Robinson *et al.*, Increased somatic mutation burdens in normal human cells due to
991 defective DNA polymerases. *Nat Genet* **53**, 1434-1442 (2021).
992 56. L. Zapata *et al.*, Negative selection in tumor genome evolution acts on essential cellular
993 functions and the immunopeptidome. *Genome Biol* **19**, 67 (2018).
994 57. L. Bányai, M. Trexler, K. Kerekes, O. Csuka, L. Patthy, Use of signals of positive and
995 negative selection to distinguish cancer genes and passenger genes. *Elife* **10**, (2021).
996 58. I. E. Vorontsov *et al.*, Negative selection maintains transcription factor binding motifs in
997 human cancer. *BMC Genomics* **17**, 395 (2016).
998 59. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in
999 141,456 humans. *Nature* **581**, 434-443 (2020).
1000 60. X. Zhang *et al.*, Genetic constraint at single amino acid resolution in protein domains
1001 improves missense variant prioritisation and gene discovery. *Genome Med* **16**, 88 (2024).
1002 61. M. Gudkov, L. Thibaut, E. Giannoulatou, Context-adjusted proportion of singletons
1003 (CAPS): a novel metric for assessing negative selection in the human genome. *NAR Genom*
1004 *Bioinform* **6**, lqae111 (2024).
1005 62. A. R. J. Lawson *et al.*, Somatic mutation and selection at population scale. *Nature* **647**,
1006 411-420 (2025).
1007 63. S. Bellusci, J. Grindley, H. Emoto, N. Itoh, B. L. Hogan, Fibroblast growth factor 10
1008 (FGF10) and branching morphogenesis in the embryonic mouse lung. *Development* **124**,
1009 4867-4878 (1997).
1010 64. S. Melboucy-Belkhir *et al.*, Forkhead Box F1 represses cell growth and inhibits COL1 and
1011 ARPC2 expression in lung fibroblasts in vitro. *Am J Physiol Lung Cell Mol Physiol* **307**,
1012 L838-847 (2014).
1013 65. N. Zoppi, R. Gardella, A. De Paepe, S. Barlati, M. Colombi, Human Fibroblasts with
1014 Mutations in COL5A1 and COL3A1 Genes Do Not Organize Collagens and Fibronectin
1015 in the Extracellular Matrix, Down-regulate α2β1 Integrin, and Recruit αvβ3 Instead of
1016 α5β1 Integrin*. *Journal of Biological Chemistry* **279**, 18157-18168 (2004).
1017 66. L. Zhang *et al.*, Single-cell whole-genome sequencing reveals the functional landscape of
1018 somatic mutations in B lymphocytes across the human lifespan. *Proc Natl Acad Sci U S A*
1019 **116**, 9014-9019 (2019).
1020 67. J. Ganz *et al.*, Contrasting somatic mutation patterns in aging human neurons and
1021 oligodendrocytes. *Cell* **187**, 1955-1970.e1923 (2024).
1022 68. T. B. L. Kirkwood, Evolution of ageing. *Nature* **270**, 301-304 (1977).
1023 69. J. Hård *et al.*, Long-read whole-genome analysis of human single cells. *Nature*
1024 *Communications* **14**, 5164 (2023).
1025 70. L. Yang, A Practical Guide for Structural Variation Detection in the Human Genome. *Curr*
1026 *Protoc Hum Genet* **107**, e103 (2020).
1027 71. P. A. Ewels *et al.*, The nf-core framework for community-curated bioinformatics pipelines.
1028 *Nat Biotechnol* **38**, 276-278 (2020).
1029 72. H. Patel *et al.*, *nf-core/rnaseq: nf-core/rnaseq v3.1.0*. (2024), vol. 3.1.0.

1030 73. T. Hart, H. K. Komori, S. LaMere, K. Podshivalova, D. R. Salomon, Finding the active
1031 genes in deep RNA-seq gene expression studies. *BMC Genomics* **14**, 778 (2013).
1032 74. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for
1033 RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
1034 75. T. Wu *et al.*, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
1035 *Innovation (Camb)* **2**, 100141 (2021).
1036 76. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
1037 *Res* **28**, 27-30 (2000).
1038 77. D. Saul *et al.*, A new gene set identifies senescent cells and predicts senescence-associated
1039 pathways across tissues. *Nat Commun* **13**, 4827 (2022).
1040 78. M. Gundry, W. Li, S. B. Maqbool, J. Vijg, Direct, genome-wide assessment of DNA
1041 mutations in single cells. *Nucleic Acids Res* **40**, 2032-2040 (2012).
1042 79. . (2015).
1043 80. F. Krueger, Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently*
1044 *apply quality and adapter trimming to FastQ files* **516**, 517 (2015).
1045 81. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-
1046 generation DNA sequencing data. *Nature Genetics* **43**, 491-498 (2011).
1047 82. M. Petljak *et al.*, Characterizing Mutational Signatures in Human Cancer Cell Lines
1048 Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294.e1220 (2019).
1049 83. L. Zhang *et al.*, Analysis of Somatic Mutations in Senescent Cells Using Single-Cell
1050 Whole-Genome Sequencing. *Aging Biol* **1**, (2023).
1051 84. F. Blokzijl, R. Janssen, R. van Boxtel, E. Cuppen, MutationalPatterns: comprehensive
1052 genome-wide analysis of mutational processes. *Genome Med* **10**, 33 (2018).
1053 85. S. Sun, Y. Wang, A. Y. Maslov, X. Dong, J. Vijg, SomaMutDB: a database of somatic
1054 mutations in normal human tissues. *Nucleic Acids Res* **50**, D1100-d1108 (2022).
1055 86. Z. Sondka *et al.*, COSMIC: a curated database of somatic variants and clinical data for
1056 cancer. *Nucleic Acids Res* **52**, D1210-d1217 (2024).
1057 87. A. Degasperi *et al.*, A practical framework and online tool for mutational signature
1058 analyses show inter-tissue variation and driver dependencies. *Nat Cancer* **1**, 249-263
1059 (2020).
1060 88. C. P. Selby, L. A. Lindsey-Boltz, W. Li, A. Sancar, Molecular Mechanisms of
1061 Transcription-Coupled Repair. *Annu Rev Biochem* **92**, 115-144 (2023).
1062 89. D. Broad. (Figshare+, 2024).
1063 90. V. A. Blomen *et al.*, Gene essentiality and synthetic lethality in haploid human cells.
1064 *Science* **350**, 1092-1096 (2015).
1065 91. M. Lawrence *et al.*, Software for computing and annotating genomic ranges. *PLoS Comput*
1066 *Biol* **9**, e1003118 (2013).

1067